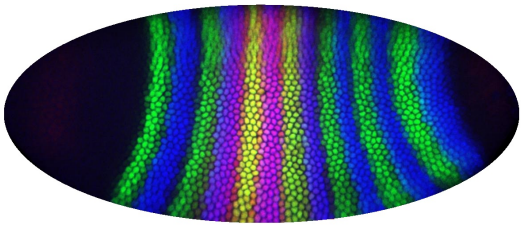
TÉCNICO LISBOA



Morphogenesis of *Drosophila* early development: the even-skipped protein

Beatriz Álvaro Soares de Albergaria

Thesis to obtain the Master of Science Degree in

Engineering Physics

Supervisor: Prof. Rui Manuel Agostinho Dilão

Examination Committee

Chairperson: Prof. Ilídio Pereira Lopes
Supervisor: Prof. Rui Manuel Agostinho Dilão
Member of the Committee: Prof. Rosa Filipa Penha Alves

September 2020

Acknowledgments

In the first place, I want to thank Prof. Rui Dilão, my supervisor in this thesis, for all his guidance, support, collaboration and very fruitful advises. In this thesis, I learned a lot about biophysics, dynamical systems and about how to conduct a research project, and I thank him immensely for that.

Concerning my education throughout these years, I want to thank everyone at Instituto Superior Técnico and at Departamento de Física for this very rewarding journey, where I had the privilege to learn a lot with great professors and colleagues.

I want to thank very much my friends who made this past time working on this thesis much more motivating and enjoyable, specially Bárbara, Cláudia, Diogo, Mariana, Nelson and Nuno. I want to thank my mother, for helping and supporting me not only during this thesis but throughout my whole life. Finally, I want to thank António for his care, support and for being my best friend.

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Resumo

A *Drosophila melanogaster* produz várias proteínas que se regulam em cascada ao longo das duas primeiras horas de desenvolvimento embrionário: enquanto que as primeiras a expressarem-se têm um perfil difuso, as últimas já exibem um padrão de alta frequência. A penúltima classe de proteínas que se expressa são as pair-rule, com um padrão de sete riscas, e que inclui as proteínas *Even-skipped* e *Fushi-tarazu*, com padrões exactamente complementares. A última classe de proteínas a estabelecer-se é a dos segment-polarity, que se identificam por um padrão de catorze riscas.

Este trabalho teve como objetivo modelar a formação da proteína *Even-skipped* através de mecanismos de reacção e de difusão. Mostrou-se que os mecanismos de reacção pelas proteínas anteriores não são suficientes para produzir um padrão de riscas. Por outro lado, ao considerar-se que os pair-rule são regulados por reacção-difusão – i.e., um mecanismo de padrões de Turing - obtiveram-se padrões transientes e um estado estacionário semelhantes aos experimentais. Para tal, utilizou-se o modelo do Brusselator para descrever o sistema. Estendendo a abordagem aos segment-polarity, foi possível simular um padrão de catorze riscas, que só se definiu totalmente quando se considerou um mecanismo de reacção-difusão, sendo as reacções locais novamente insuficientes.

De modo a determinar o sistema dinâmico desconhecido que regula estes genes, utilizou-se um algoritmo de regressão ao estado estacionário, que identificou os termos de interacção. No entanto, esta solução divergiu quando integrada, o que significa que é necessária uma melhor caracterização da validade das soluções deste algoritmo.

Palavras Chave

Drosophila melanogaster, segmentos, *pair-rule*, *segment-polarity*, reacção-difusão, padrão de Turing

Abstract

Drosophila melanogaster produces several proteins throughout its first two hours of embryonic development, which are regulated in a cascade manner: the first have an ill-defined profile and the last to appear have a high-frequency pattern. The second to last of these embryonic proteins to be expressed are the pair-rule proteins, with a seven-stripes pattern, and these include *Even-skipped* and *Fushi-tarazu*, which have perfectly complementary patterns. The last proteins to be expressed are the segment-polarity, with a pattern of fourteen stripes.

The purpose of this work was to model the development of the *Even-skipped* protein considering both reaction and diffusion mechanisms. We showed that reaction mechanisms, controlled by the earlier proteins, are not sufficient to produce a striped pattern. Nevertheless, when we considered an interaction between the pair-rule as well as diffusion processes – that is, a Turing pattern mechanism - we obtained transient patterns and a steady-state that closely resemble the experimental data. For this, we used the Brusselator model to describe the system. Thereafter, we extended this approach for the segment-polarity genes and we were able to obtain a fourteen-stripes pattern, which only became completely defined when we considered a reaction-diffusion mechanism, and the local reactions were once more insufficient.

In order to determine the unknown dynamical system that regulates these genes, we used a regression algorithm that clearly identified the interaction terms. Nonetheless, the solution diverged when integrated, which means that the validity of the algorithm's solutions needs to be better characterized.

Keywords

Drosophila melanogaster, segments, pair-rule, segment-polarity, reaction-diffusion, Turing patterns

Contents

List of Figures

List of Tables

1

Introduction

Contents

1.1 Drosophila early development

In insects, the definition of periodic band structures along the antero-posterior axis of the body occurs early in embryo development. This periodic structure determines the segment organisation of the embryo, conditioning the following morphogenic processes. These processes are going to set the differentiation of adult functional areas that follow the insect development.

For this periodic structure to emerge, just like all morphological structures, several complex networks of gene regulatory pathways need to be developed at the cellular level. *Drosophila malanogaster*, also known as the *fruit fly*, is a commonly used insect in experimental biology due to its simple genetic structure. Due to this, it has a large amount of experimental data sources. It is, therefore, a good choice for the object of study if one aims to make a mathematical model of the gene regulation mechanisms [1].

The first coordinates of *Drosophila*'s shape are set before fertilisation. The mother places mRNA at the poles of the oocyte: a concentration of caudal (*bcd*) mRNA is set at the future anterior pole, and a concentration of *cad* is set at the future posterior pole. Thus, when the mature egg is laid, it is both morphologically and molecularly polarised.

After fertilisation, the maternal genes are translated into proteins. In this stage, both *bcd* and *cad* mRNAs are transcribed into *Bcd* and *Cad* proteins, that diffuse away from the poles and become gradually distributed over about one-half of the length of the egg - see figure 1.1 for the evolution of their profile. These maternal proteins are going to determine large body parts: *Bcd* defines the anterior axis while *Cad* defines the posterior axis [1] [2]. The first 13 nuclear divisions (that are not followed by cytoplasm division) give rise to a *syncytial blastoderm*, in which all the nuclei share the same cytoplasm and have migrated along the cell's surface. It is, therefore, very easy for morphogenes and nutrients to spread between nuclei [1]. During this stage, the *zygotic genes* are expressed: these are responsible for the definition of large body parts, just like the maternal genes, but are also responsible for progressively differentiation in the fly's body pattern.

The zygotic genes are transcribed in certain regions of the embryo's syncytial blastoderm, and the resulting proteins will act as transcription factors that regulate genes that will be transcribed afterwards [1]. Accordingly, we observe a hierarchy, or cascade, of genes: the genes to be expressed earlier regulate the activity of those to be expressed later, where the former control large domains - entire body regions - and the later control minor domains, that result in segmentation patterns [2] [1]. The first zygotic genes to be expressed the are the *gap genes*, that affect entire regions of the body, and these are regulated by maternal-effect and gap proteins. These gap genes include giant *(gt)*, kruppel *(kr)*, hunchback *(hb)* and knirps *(kni)* - for their concentration evolution, see figures 1.1 and 1.2. These plots were made with fluorescence images from the FlyEx database, and the concentration profiles for the different proteins were normalized to *Eve* maximum at cycle 14A8; we used the central part of the embryo, and the noise was filtered. After that, the expression of *pair-rule genes* occurs, which are the first proteins to show a periodic structure,creating a seven band structure perpendicularly to the antero-posterior axis; these include fushi-tarazu *(ftz)*, even-skipped (*eve*), runt *(run)*, sloppy-paired

(slp), odd-paired *(opa)*, odd-skipped *(odd)*, hairy (*h*) and paired *(prd)* - see figure 1.3 for the evolution of the concentration profile of the first two, together with maternal *bcd* and *cad*. In figure 1.4, we show the pair-rule and the gap genes profile at stages 14A1 and 14A8 for comparison.

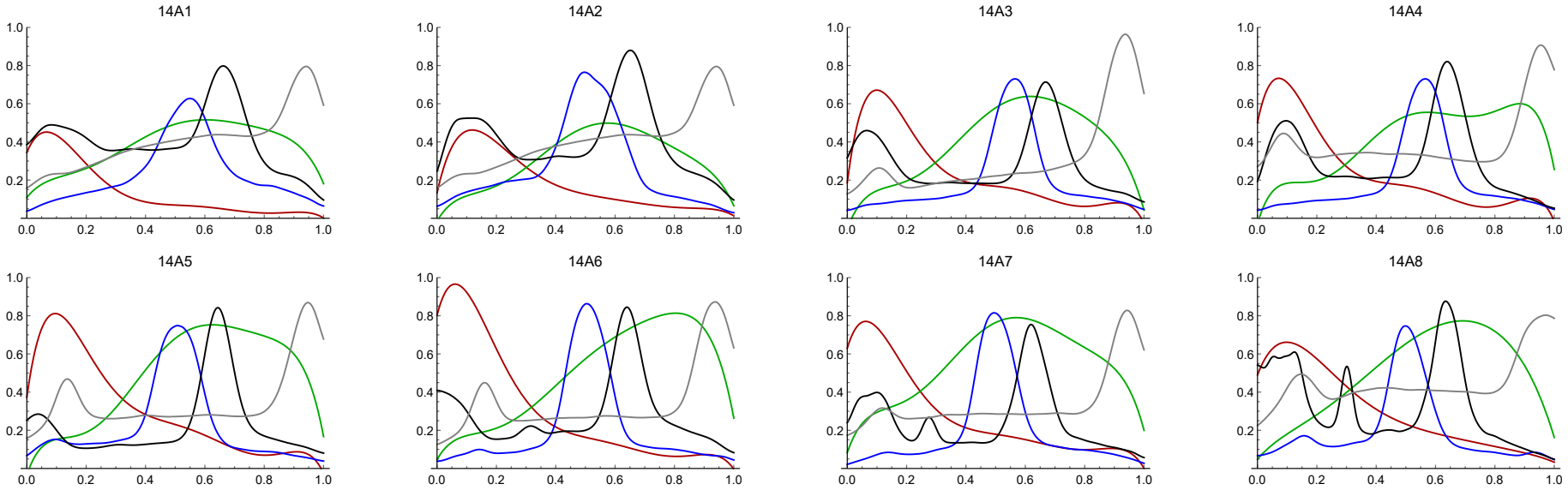


Figure 1.1: Normalized concentration through normalized embryo length of gap proteins *Bcd* (red), *Cad* (green), *Kr* (blue) and *Kni* (black) in relation to *Eve* concentration. These data, shown in several phases of 14th cycle, was taken from the FlyEx database and was convolved with a gaussian filter. Code at *export data to excel.nb* - this code file, as all the others in this thesis, are available from the corresponding authors, *beatriz.albergaria@tecnico.ulisboa.pt* and *ruidilao@tecnico.ulisboa.pt*, on reasonable request.

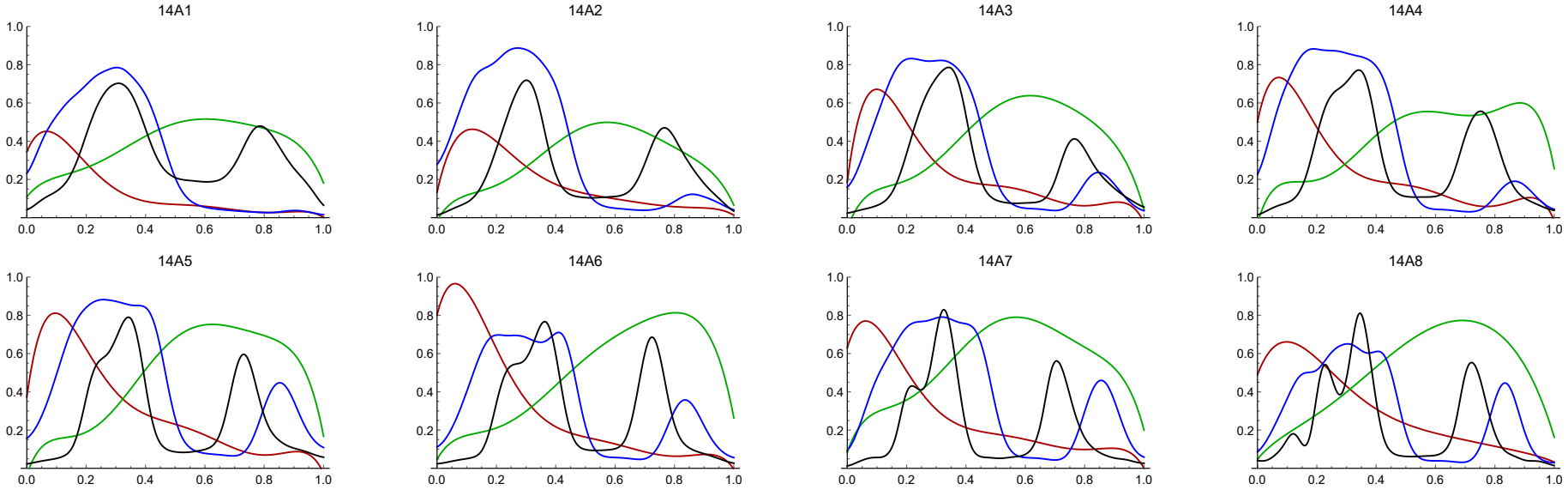


Figure 1.2: Normalized concentration through normalized embryo length of gap proteins *Bcd* (red), *Cad* (green), *Gt* (blue) and *Hb* (black) in relation to *Eve* concentration. These data, shown in several phases of the 14th cycle, was taken from the FlyEx database and was convolved with a gaussian filter. Code at *export data to excel.nb*.

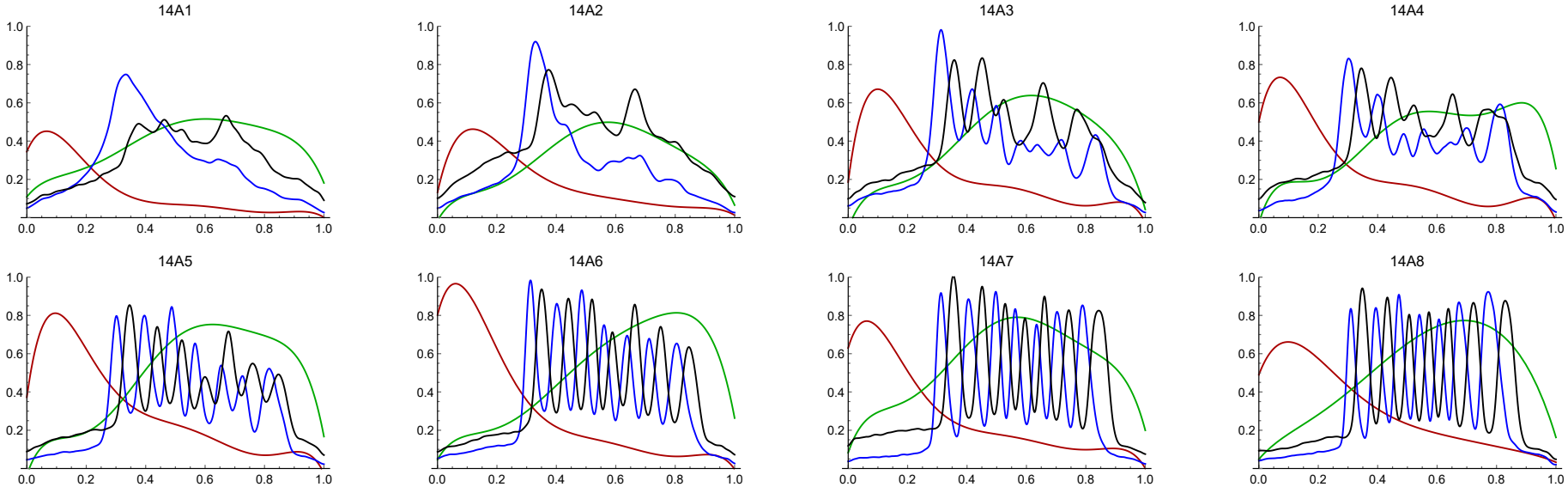


Figure 1.3: Normalized concentration through normalized embryo length of maternal proteins *Bcd* (red), *Cad* (green), *Eve* (blue) and *Ftz* (black) in relation to *Eve* concentration. These data, shown in several phases of the 14th cycle, was taken from the FlyEx database and was convolved with a gaussian filter. Code at *export data to excel.nb*.

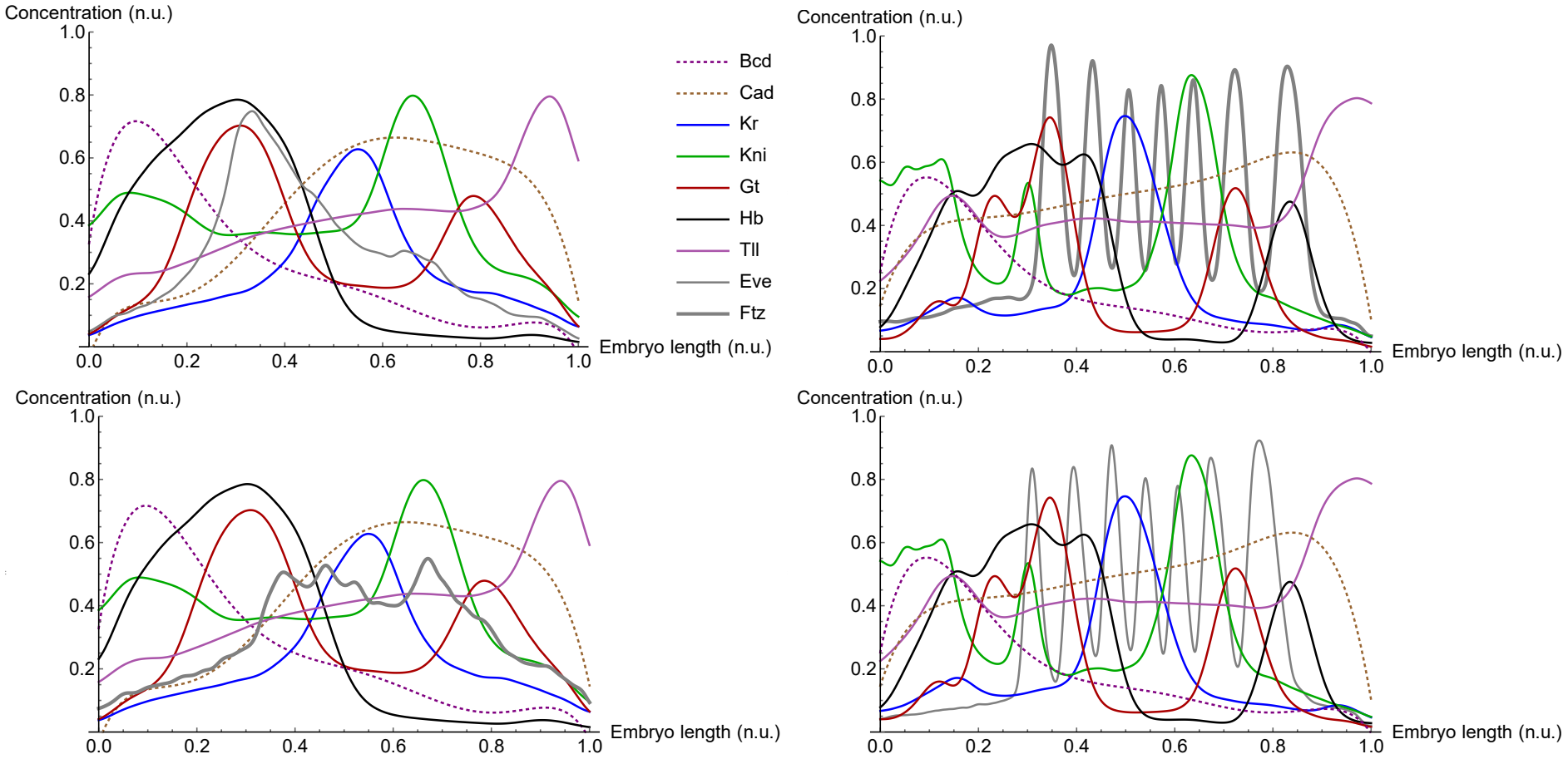


Figure 1.4: Gap, maternal and pair-rule profiles at cycle 14A1 (left top and bottom) and at cycle 14A8 (right top and bottom). Code at *export data to excel.nb*.

Finally, the expression of the *segment-polarity genes* takes place: these genes affect every single segment and are responsible for the doubling of the fly's stripes from seven to fourteen; these include the engrailed *(en)*, wingless *(wg)*, gooseberry *(gsb)*, patched *(ptc)* and hedgehog *(hh)* genes [2] [3] - see figure 1.5 for the pattern of *en* as an example.

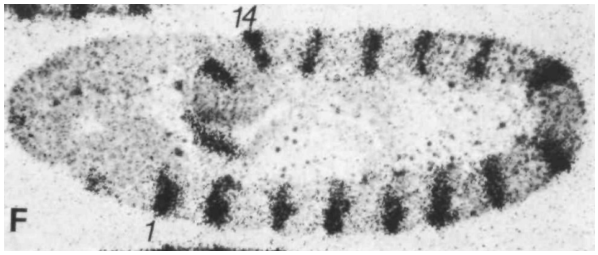


Figure 1.5: Adjacent sections of an embryo approximately 2h after gastrulation hybridized with 35S-probes for *en*. Taken from figure 2F at [4].

At this stage - interphase of the 14th mitotic cycle, approximately 2 hours after fertilization - occurs *cellularization*: the cytoplasmic membranes begin to form around the nuclei of the syncytial blastoderm, the individual cells become allocated into one of the of the fourteen visible segments giving rise to a *cellular blastoderm* [1] [2]. In figure 1.6 we show a summary of these development stages of *Drosophila*'s genetic expression.

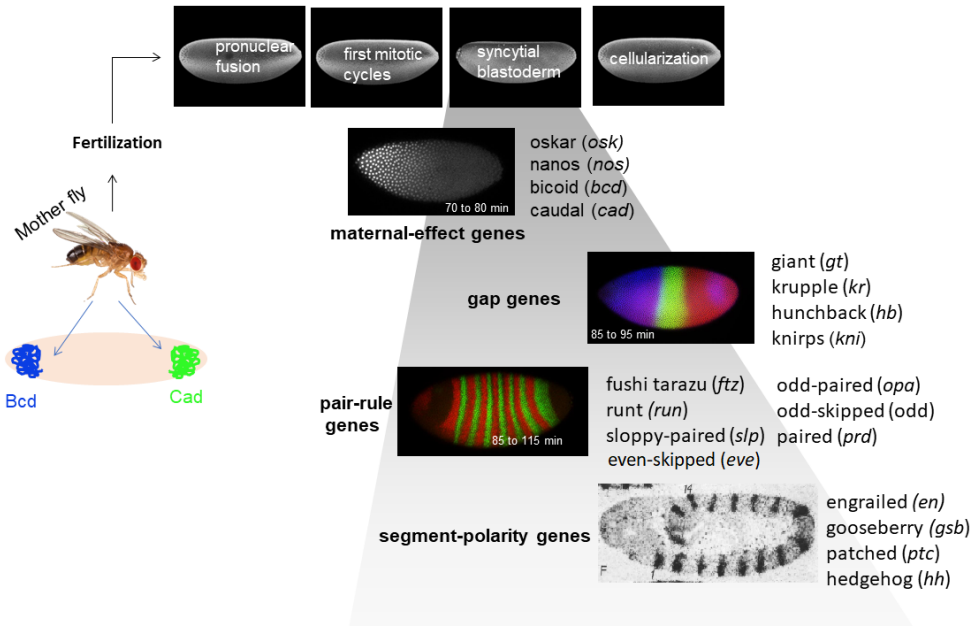


Figure 1.6: Summary of *Drosophila* development stages and the genetic expression that occurs in each. Images from the FlyEx database and figure adapted from [5].

1.2 Bibliographic review on stripe formation

1.2.1 Principle stripe mechanism

The majority of experiments made in order to understand pair-rule's stripe formation have a biological approach, and are based on the *principle stripe mechanism*. According to this view, the

pair-rule gene expression is controlled in transcription, such that transcript levels are enhanced within stripe domains and diminished between them, that is, the stripe region itself is activated whereas the interstripe region is repressed. Taking into account that the *Drosophila* genes are expressed in a cascade manner, the pair-rule genes are thought to respond directly to gap gene positional cues via extensive upstream promoters with independent regulatory elements for individual stripes. In other words, the stripes are thought to be formed individually - each of them is activated and repressed in its boundaries by a specific set of genes [6].

1.2.2 *Eve* regulation

In order to understand how the gradient concentration of maternal-effect, gap and pair rule genes define the periodic structure of seven bands that characterize *eve* and *ftz* expression, several experiments have been made inhibiting one gene expression and checking how the *Eve* pattern is modified.

The maternal products that are reported to have the greater impact in the evolution of *Eve* pattern are the *Bcd* and the *Cad* proteins. Conversely, the more important gap genes are *gt*, *kr*, *hb* and *kni* [7]. We shall now analyze what is the role that each of these maternal and zygotic proteins has in each of the seven stripes by analyzing some of these experiments.

1.2.2.A Maternal proteins regulation

Bcd protein

Firstly, in what concerns experiences with maternal genes, we start by analyzing the experiences made with *bcd* mutation. In [8], stripes 3 and 7 were stained for the control *eve* pattern (with no muations) and for the embryo with *bcd-*: in this paper, it is reported an enormous reduction of stripe 7 intensity and an anterior expansion and shift of stripe 3 – see figures 1.7 and 1.8 for the reproduction of these results.

Figure 1.7: Control pattern: wild-type, cellularizing embryo hybridized with an *eve* antisense RNA probe. The staining pattern consists of seven transverse stripes along the antero-posterior axis. Taken from figure 3 A of [8].

Figure 1.8: Mutant lacking *bcd*+ function. Cellularized embryo carrying the 500-bp *eve* stripe 3/7 enhancer attached to the *eve*–lacZ fusion gene. Taken from figure 3 C of [8].

Moreover, in [9], besides stripe 3 and 7, stripe 2 was investigated as well. Firstly, the *bcd* binding site for stripe 2 (bcd2) was inhibited and there was not a repported change in stripe 2 (results not

shown in the paper). The inhibition of *kr* binding site for stripe 3 (kr3) and *bcd* binding site for stripe 1 – which the authors claim to be the same – did not cause a discernible effect in stripe 2 as well. On the contrary, the inhibition of both binding sites bcd2 and kr3/bcd1 resulted in a decreasing staining of stripe 2 and no alterations for stripes 3 and 7 – see figures 1.9 and 1.10.

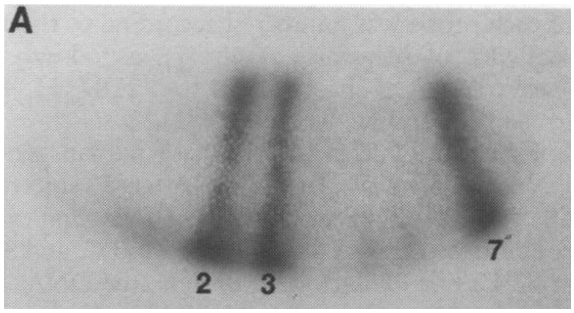


Figure 1.9: Control pattern: P-transformed hybridyzed embryo expressing the wild-type 5.2-kb *eve*-lacZ fusion gene. Taken from figure 2 A in [9].

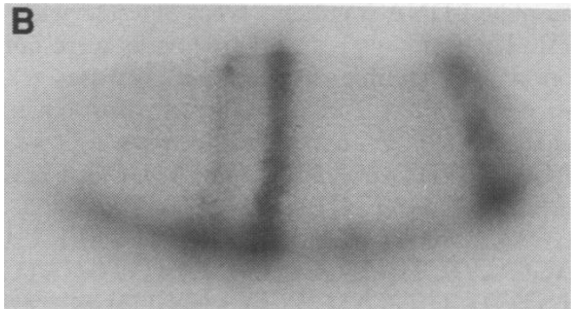


Figure 1.10: Pattern obtained with the double mutant in the bcdl-kr3 sequence and the bcd2 site. Taken from figure 2 B in [9].

1.2.2.B Gap proteins regulation

Kr protein

We now move to the experiences concerning mutations with gap genes. In mutants lacking *kr* function, it is reported in [9] that there is a more pronounced expression of stripe 2, when stripes 2, 3 and 7 are stained – compare figures 1.11 and 1.12. The very same result for stripe 2 is reported in [10].

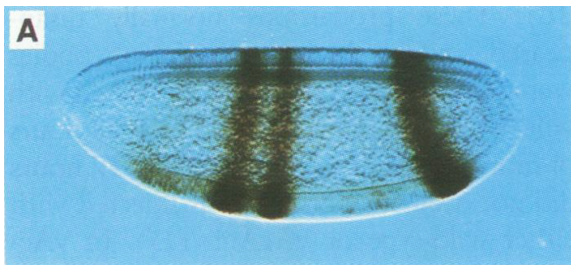


Figure 1.11: P-transformed embryo expressing the wild-type 5.2-kb *eve*-lacZ fusion gene. Stripes 2, 3, and 7 are stained to the same extent. Taken from figure 3 A in [9].

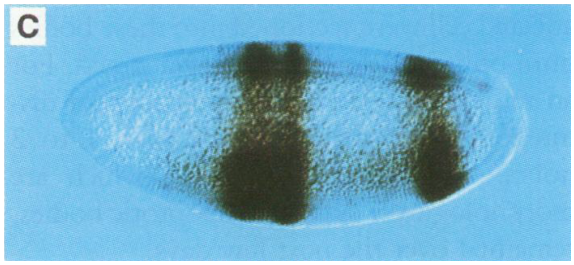


Figure 1.12: Expression of the *eve*-lacZ fusion gene containing point mutations in all six high-affinity *Kr* protein binding sites. There is a more pronounced expansion of stripe 2. Taking from figure 3 C in [9].

These results are confirmed by [11], in which the authors followd *eve* pattern (in time) in *Kr*- embryos. Indeed, if we compare the *Kr* mutant with the wild-type, we see that the former has a stripe 2 which is enlarged as it is extended to the posterior part of the body. Moreover, the structures of stripes 4, 5, 6 and 7 are transformed into two new different stripes. These results are reproduced in

figures 1.13 and 1.14, and in 1.16 we make an overlap of the final patterns of both *Kr*- and wild-type for a more clear comparison.

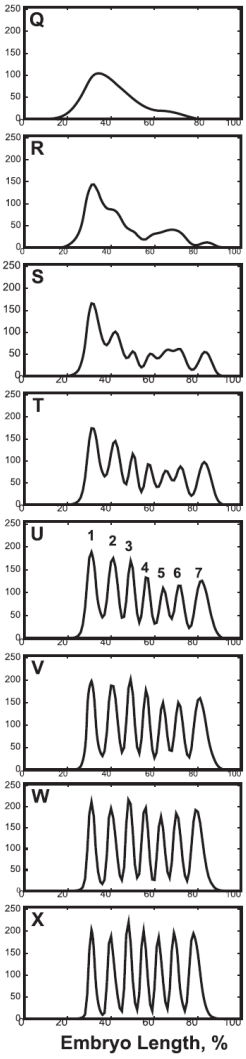


Figure 1.13: Control: one-dimensional *eve* integrated pattern in 8 time classes of cycle 14 (6.5 min long each). Taken from figures 1 Q-X in [11].

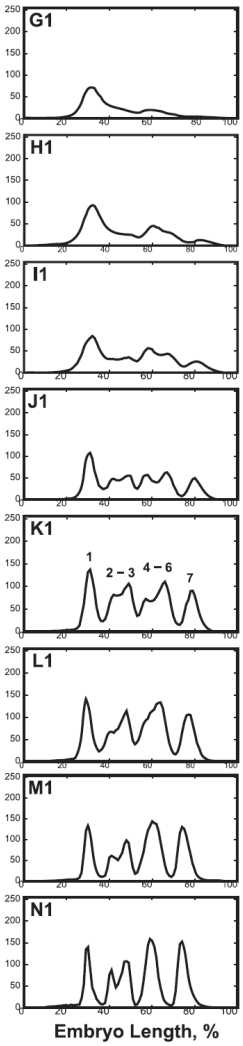


Figure 1.14: *eve* expression in *Kr*- embryo. Taken from figures 1 G1-N1 in [11].

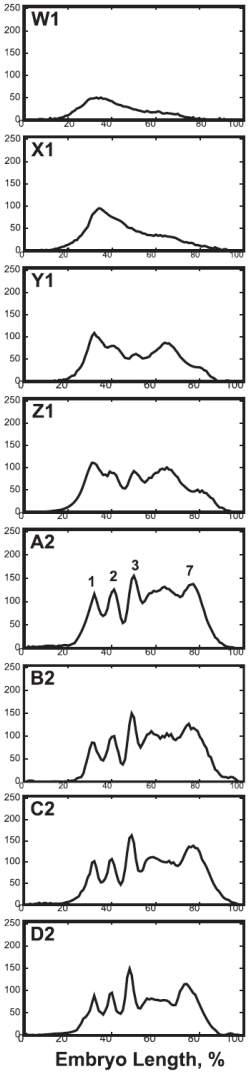


Figure 1.15: *eve* expression in *kni*- embryo. Taken from figures 1 W1-D2 in [11].

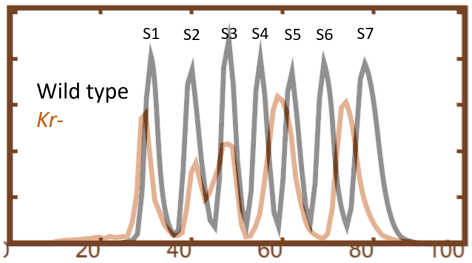


Figure 1.16: Overlap of the expression of *eve* in wild-type (figure 1 X in [11]) and *Kr*- embryo (figure 1 N1 in [11]) at stage 14.

Kni protein

In what concerns the gap gene *kni*, it is reported in [8] that the *eve* pattern shows significant alterations in the posterior part of its mutants: although stripes 1 and 2 remain normal, there is a broad staining that encompasses stripes 3 through 7 – this result is replicated in figure 1.17.

Figure 1.17: *eve* staining pattern in a *kni* mutant at the midpoint of nuclear cleavage cycle 14. This staining pattern was visualized by *in situ* hybridization using an *eve* anti-sense RNA probe. The wild-type expression in not shown at this stage. Taken from figure 5 A in [8].

Once again, this result is supported in [11] with experiences made with *kni-*, analogous to the ones described before for *Kr-*. In figure 1.15, we see that *eve* pattern in *kni-* mutants initially develops as in the wild-type but, in later stages, the most posterior stripes are unable to form; in figure 1.18 we show the overlap of the final patterns of both *kni-* and wild-type for a more clear comparison.

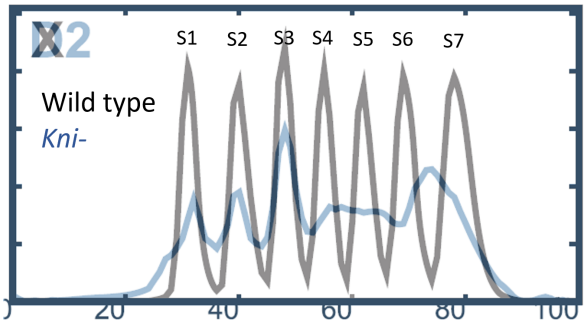


Figure 1.18: Overlap of the expression of *eve* in wild-type (figure 1 X1 in [11]) and *kni-* embryo (figure 1 D2 in [11]) at the last time class of cycle 14.

In a later stage of development, the cellularizing embryo, it is reported in [8] that this *eve* pattern (in *kni-* mutants) persists, nevertheless, the intensity of staining which covers stripes 3-7 is reduced, so that the posterior border of stripe 3 has formed.

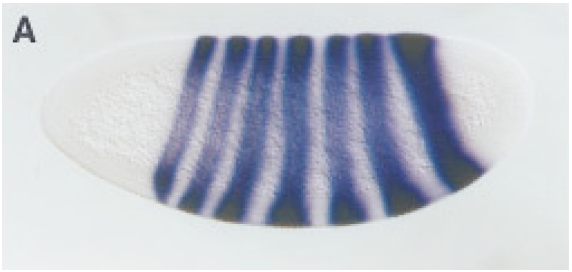


Figure 1.19: Wild-type cellularizing embryo visualized by *in situ* hybridization using an *eve* anti-sense RNA probe. Taken from figure 2 A in [8].

Figure 1.20: *eve* pattern in a *kni-* mutant which has completed cellularization – same as figure 1.17 but the embryo is 20-30 min older. Taken from figure 5 B in [8].

***Hb* mutants**

In [8] it is also presented the stripes 3 and 7 of *eve* in *hb* mutants, where it is observed that the anterior border stripe 3 is expanded (toward the anterior region), whereas the posterior border of stripe 7 is also extended (toward the posterior region); moreover, the staining of the latter stripe is intensified. We can roughly observe that the stripe 3 extended stain covers stripe 2 and perhaps 1 as well, nevertheless, this is not a very precise conclusion since the paper does not present the *eve* control pattern for this stage of development.

Figure 1.21: Staining pattern in mutants lacking the zygotic component of *hb* gene activity. Stage is not specified but, from the sequence, we assume it is a cellularizing embryo. Taken from figure 4 C in [8].

These results are not in agreement to what is reported in [12], where there is a complete or partial deletion of parasegments from 2 to 7 (which include *eve* stripes 2, 3 and 4) as well as a reduction in stripe 7 intensity, in both cellular blastoderm stage and gastrulating embryos.

Recalling the results shown for *kni* and *hb* mutants, we saw that the absence of the former gene resulted in a broad staining which encompassed stripes 3 to 7 (and no alterations for stripes 1 and 2), while mutations in the latter gene were followed by an extended stain in stripe 3 towards the anterior part and in stripe 7 towards the posterior region. This way, we can predict that the inhibition of both *hb* and *kni* transforms *eve* striped pattern into a broad stain. In fact, [12] reports that, in *hb*- and *kni*- cellular blastoderm-stage embryos, *eve* and *ftz* proteins are not longer distributed within stripes but form a continuous staining between 13% and 74% of the embryo length - see figures 1.23 and 1.24 respectively, figure 1.22 for control staining at the same stage -, with the exception of a few discontinuities (arrows in the figures).

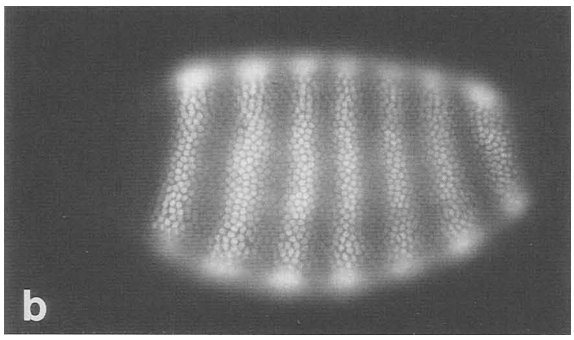


Figure 1.22: Cellular blastoderm-stage wild-type embryo after staining with anti-*eve*. Taken from figure 1 b in [12].

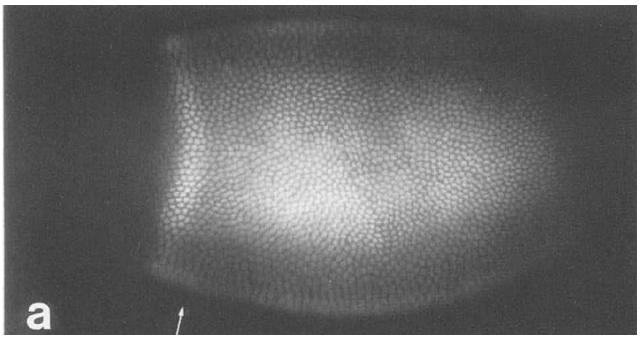


Figure 1.23: Cellular blastoderm-stage double mutant stained with anti-*eve*. Taken from figure 1 a in [12].

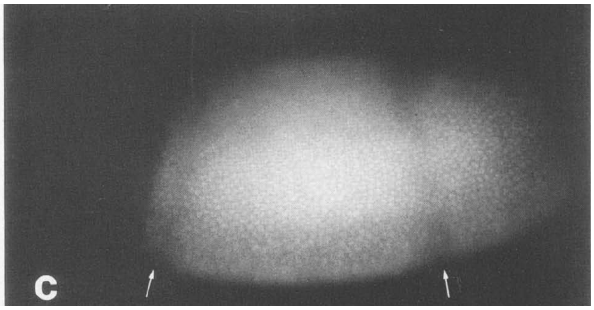


Figure 1.24: Cellular blastoderm-stage double mutant stained with anti-*ftz*. Taken from figure 1 c in [12]

Gt mutants

In what concerns *gt* mutants, we see from the experiment made in [9] that without the expression of this gap gene the anterior border of *eve*'s stripe 2 becomes substantially extended to the anterior part – see figure 1.9 for the 2, 3 and 7 stripe control and figure 1.25 for the same staining but in *gt* mutants.

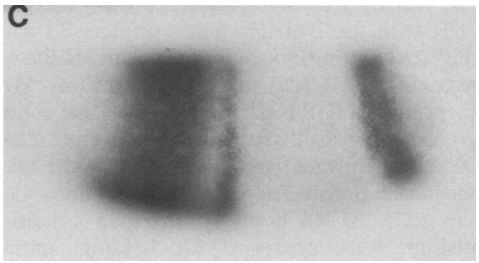


Figure 1.25: The expression pattern from an *eve*-lacZ fusion gene containing two small deletions that remove all three *Gt* protein binding sites. Taken from figure 2 C in [9].

The very same result was obtained in [10], where only stripe 2 was tracked - see figure 1.26 for the control and 1.27 for the *gt* mutant.

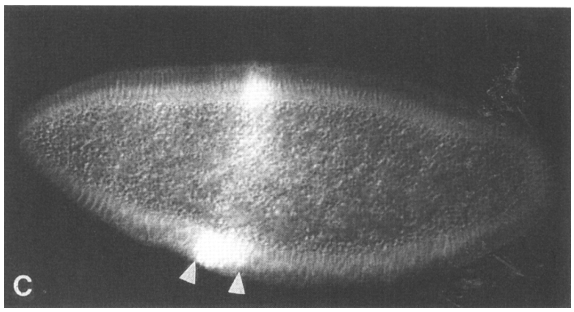


Figure 1.26: Expression of the lacZ reporter gene was detected by staining with an anti-β-galactosidase antibody. Staining is restricted to stripe 2; none of the other *eve* stripes are observed including 7. Taken from figure 2 C in [10].

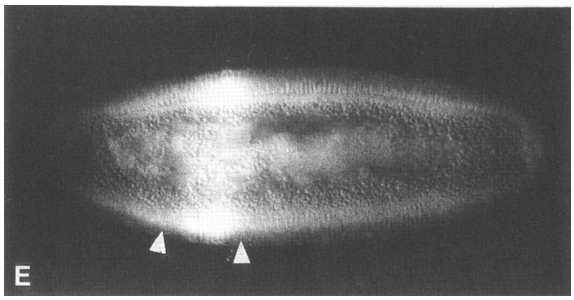


Figure 1.27: Stripe 2 expression in a gt- embryo. The eve-lacZ fusion gene shown in B was crossed into a gt^{YAs2} mutant background. Taken from figure 2 E in [10].

In [12], there is additional information for cellular blastoderm stage embryos lacking *gt* function, as all stripes were tracked: firstly, we come to know that the anterior expansion of stripe 2 is in fact a composite band of stripes 1 and 2; furthermore, in *gt* mutants, *eve*'s stripes 5 and 7 are merged into a broad staining; finally, in the mutant embryo there is a reduced intensity of stripe 7 expression. In figure 1.28, we see the staining pattern which results from this mutation, and we recall figure 1.22 for the control pattern.

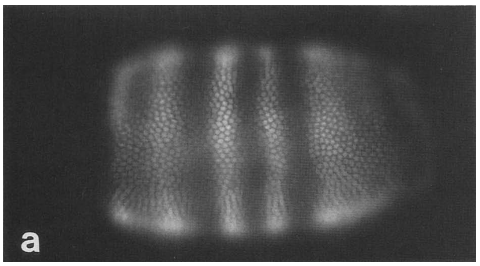


Figure 1.28: Cellular blastoderm-stage embryo after staining with anti-*eve*. Taken from figure 7 A in [12].

Pair-rule regulation

Moving to the pair-rule genes, it is reported in [12] that *prd*, *odd*, *opa* and *slp* do not appear to affect neither the establishment nor the maintenance of the *eve* pattern (results not shown in the paper). Yet, *h* and *run* are essential for the normal maintenance of the *eve* expression during gastrulation - no abnormalities during our stage of study, *i.e.*, until the end of syncytial blastoderm stage, are reported [12]. In *run*- embryos, the first alteration is detected after cellularization, where stripe 5 is narrower and has a less intense staining than in wild-type. However, on the onset of gastrulation, *eve* stripes (except 5) broadens an average of five cells. In *h* mutants, on the contrary, there is an overall decreasing of intensity and narrowing of stripes from the end of cellularization until gastrulation, and only weak staining of stripes 1, 3 and is observed at the beginning of long germ band elongation. In summary, in *run*- gastrulating embryos, *eve* pattern is over-expressed and *ftz* products disappear prematurely; in contrast, in *h*- the *eve* expression fades prematurely while *ftz* is overexpressed [12]. We do not reproduce this results here since, as we mentioned, these stages are not very relevant for the development period we are studying, nonetheless these results can be consulted in figures 9 A, B, C and D of [12]. We should note that, in all these mutants (gap and pair-rule), whatever is the alteration in *eve* pattern, a complementary expression of *ftz* is always observed [12].

In what respects *eve* itself, genetic studies also suggest that this gene contains auto-regulatory elements [13] [14]. In what concerns *eve* and *ftz*, it is mentioned in [12] that *ftz* does not appear to affect the establishment or the maintenance of the *eve* pattern, and these results are not shown in the paper. On the contrary, it is reported in [15] that, although *ftz* pattern initiation and evolution does not require *eve* function, its maintenance and refinement does.

More specifically, in early stage 14, the *eve*- embryos presents four broad bands of *ftz* expression, while in the wild-type embryo the anterior-most band is half the width of the three more posterior bands, as we can see in figure 1.29. After that, at mid-point of stage 16, both *eve*- and wild-type embryos have seven equally spaced *ftz* stripes; nonetheless, in *eve*- embryos, the anterior-most

stripe shows reduced labeling after the stripes are established, as we show in figure 1.30. At the cellularization stage, the mutating embryos show a further reduction of *ftz* expression in the anterior part, when comparing to the wild-type pattern - see figure 1.31. After cellularization is complete, the *ftz* pattern in *eve*- embryos shows particular abnormalities and irregularities in width and spacing of stripes, and the overall *ftz* expression in *eve*- is less stable than in wild-type. Finally, in gastrulation stage, the *ftz* stripes persist in wild-type embryos, while in *eve*- mutants the stripes rapidly diminish, such that by the onset of germ band elongation there is, at least, two to fourfold reduction (comparing with wild-type), as we show in figure 1.32 [15]. The corresponding figures for the *ftz* expression in wild-type embryos (at the same stages and in the same sections) are not presented in the paper.

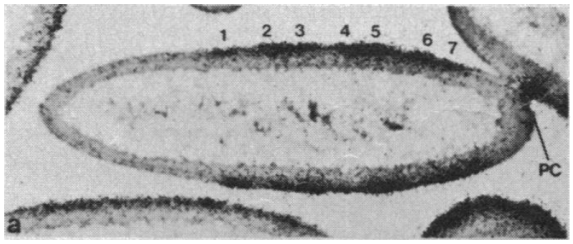


Figure 1.29: $Df(2R)eve^{1.27}$ homozygotes ($eve-$) after hybridization with the *ftz* probe (figures 1.29-1.32). Horizontal section through an early stage 14 *eve*- embryo. Taken from figure 4 a in [15].

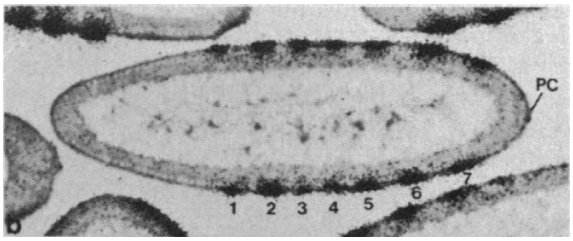


Figure 1.30: Horizontal section through an *eve*- embryo at the midpoint of stage 14 of development. Taken from figure 4 b in [15].

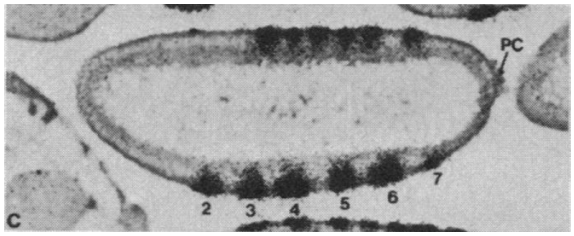


Figure 1.31: Sagittal section through a late stage 14 *eve*- embryo undergoing cellularization. Taken from figure 4 c in [15].

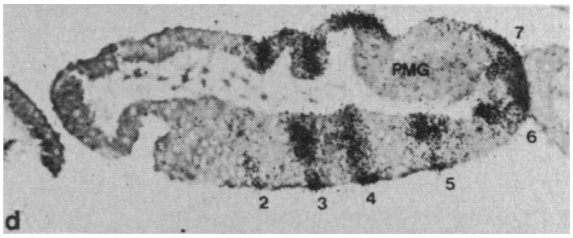


Figure 1.32: Sagittal section through an *eve*- embryo at the start of germ band elongation. Taken from figure 4 d in [15].

1.2.3 *Ftz* regulation

Unlike *eve*, *ftz* regulation is typically studied with the expression of other pair-rules. *Eve* is commonly classified as a *primary pair-rule*, together with *h* and *runt*, since its expression is determined by gap and maternal genes that are regionally distributed. These primary genes are then going to act as a mediator: they are going to translate non-periodic information of gap and maternal proteins into periodic information. Conversely, *ftz* and *prd* are classified as *secondary pair-rule* genes, as their pattern of expression is thought to be regulated directly by the primary pair-rule, and are responsible for the further establishment of the segment-polarity genes pattern [16]. This classification was first proposed by Philip William Ingham and Alfonso Martinez-Arias in [17], but there are evidences that, although this hierarchically-acting pair-rule mechanism may be conceptually appealing, it is not likely

that it is what happens in *Drosophila*'s pattern formation [16]. In [16], the *ftz* pattern was observed as mutations in the primary pair-rule *eve*, *run* and *h* were separately made, and the conclusion was barely the same for the three of them: none of these genes is key for *ftz* pattern formation, as the stripes' initial elaboration is not affected by the mutations, nonetheless, the pattern fades faster when the mentioned pair-rule are not present. Concerning *h*, a contradictory information is presented in [18], that reports an enlargement of *ftz* stripes in early stages of development in the absence of the former gene.

The *ftz* gene has also been studied on the genetic level, and its regulation is reported to be controlled by two *cis-acting elements* (that is, regions of non-coding DNA which regulate the transcription of neighboring genes [19]): the *zebra element*, which confers the striped pattern by mediating the effects of other genes, as the pair-rule mentioned above, and the *upstream element*, an enhancer element requiring *ftz* activity for its auto-regulatory action [20]. Concerning the latter, [20] states that these sequence for positive auto-regulatory feedback mechanism are only important for the enhancing and maintenance of the pattern, wheres its initial formation is set by the interaction of the pair-rule with the zebra element. In summary, there are evidences that the pair-rules *ftz* and *eve* are regulated by other genes as well as by themselves, but no pair-rule has shown to be fundamental for their development.

1.3 Motivation and objectives

The maternal and the gap genes have been analytically modeled and calibrated, considering their genetic regulatory network [21] [22]. For the *even-skipped* gene, several proposals have been made for the complete regulatory network that underlines *eve* development, taking into account the mutation experiments described above, and according to the principle stripe mechanism. Via statistical relationships between the stripes positions and the concentration profiles of gap and maternal genes, each of *Eve* stripe's expression has been successfully predicted, and a group of repressors and activators for each has been identified [23] [24]. Enhancers and promoters for stripes have also been identified [13] [25], but it is not clear that each stripe has an individual set of promoters.

The objective of this thesis is, therefore, to model both the pair-rules steady-state profiles and their evolution in time without recurring to the principle stripe mechanism. In chapter 2, we propose a model which considers activation and repression of *Eve* by the earlier stage proteins, since this is the approach considered in the majority of the literature. In chapter 3, we move to test a reaction-diffusion model, where *Eve* pattern is produced by a Turing mechanism. Finally, in chapter 4, we propose a dynamical system identification algorithm to accurately identify the model equations.

2

Reaction only model

Contents

2.1 Basics of protein synthesis

The central dogma of molecular biology is a well-established theory for the transmission of information in living organisms, originally presented by James Watson and Francis Crick: DNA is copied to other DNA molecules in cell division in *replication*, and it is also copied to an mRNA molecule via *transcription* in order to form a template for protein synthesis through *translation* [26].

We now focus on how activation and repression mechanisms may happen in gene expression. In the original DNA molecule there is a *binding site* for a specific gene, to which a *transcription factor* (TF), which is a protein, may bind. There are DNA sequences, the *enhancers*, which are binding sites for regulatory proteins that affect RNA polymerase activity. If the TF is a *promoter*, the RNA polymerase is free to bind to the binding site and initiate the transcription process; if the TF is a *repressor*, the RNA polymerase is unable to connect and gene expression does not occur. There is also the possibility that the protein which is translated is the promotor or the repressor of its own transcription, and in these situations we are dealing with self-activation or auto-catalysis and self-repression, respectively [27][28]. A scheme of these regulatory activities can be found in figure 2.1.

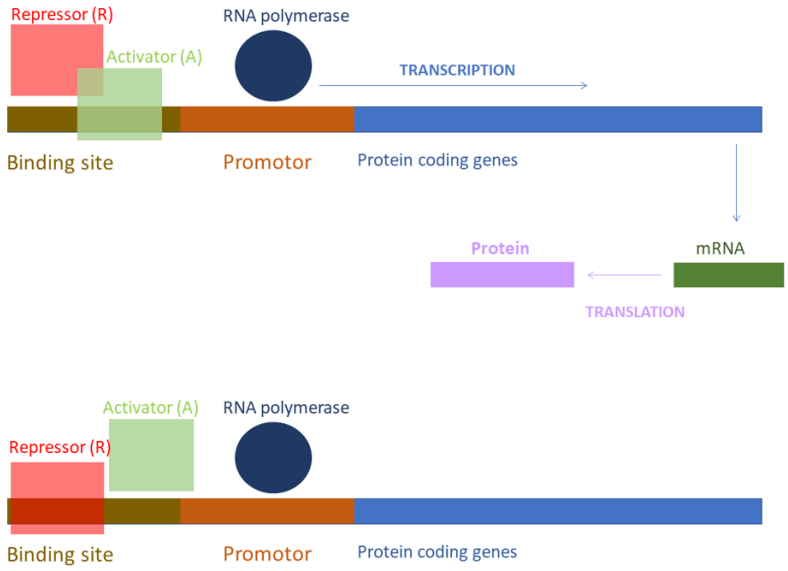


Figure 2.1: Description of the activation and repression mechanisms.

2.2 The reaction mechanism - law of mass action

In order to study chemical and biological processes in a quantitative manner, we need to describe the kinetics of the reactions that are involved: that is, given the individual reactions that are likely to occur between all possible chemical species in the system, if we want to model the rates at which the system's species concentrations evolve with time, and how these rates depend upon those concentrations, we then need to write the *reaction rate equations* [29]. These reactions are the elementary steps that make a possibly more complex reaction, being envisaged as occurring in a single collision

between the reactants. The rate at which these reactions occur is k, the *reaction rate constant*: normally, these rates are independent of concentration but are not independent of temperature, but we will not analyze the latter case. To derive these equations, we assume that in this system [30]:

- substances have low density and are homogeneous in solution;
- all reactions occur at constant pressure and temperature;
- the individual motion of the molecules in the media is independent from the other molecules, behaving as Brownian particles, and their collision frequency is proportional to the local concentrations;
- while reactions occur, the instantaneous densities do not vary.

In what concerns the formalism of a chemical reaction, the system we consider has a total number of *m* chemical substances and *n* chemical reactions. Representing the species labeled j by A_j, the reaction rate of the i^{th} reaction by k_i and the stoichiometric coefficients of the reactants and products by ν_{ij} and μ_{ij} respectively, the reactions occurring in the media can be represented by *n* collision diagrams [30]:

$$\nu_{i1}A_1 + \ldots + \nu_{im}A_m \rightarrow^{k_i} \mu_{i1}A_1 + \ldots + \mu_{im}A_m \ (i = 1, ..., n). \tag{2.1}$$

Furthermore, if $\nu_{ij} = \mu_{ij} > 0$, the corresponding substance A_j is a *catalyst* and, if $\mu_{ij} > \nu_{ij} > 0$, A_j is an *autocatalyst* [31]. Under these conditions, the time evolution of the concentration of all the chemical substances is described by the set of m ordinary differential equations - the *law of mass action* [30]:

$$\frac{dA_j}{dt} = \sum_{i=1}^{n} k_i(\mu_{ij} - \nu_{ij})A_1^{\nu_{i1}} \ldots A_m^{\nu_{im}}. \tag{2.2}$$

From these equations, we see that the modulus of the rate of time evolution of A_j will be greater the greater is the reaction rates k_i (that is, the faster is the reaction), and the greater is the difference of this species' stoichiometric coefficients as reactant - ν_{ij} - and product - μ_{ij} - (that is, the greater is the amount of A_j created or lost).

If there is mass conservation, we can also write the *conservation equations*:

$$\sum_{j=1}^{n} \alpha_{jk}A_j = \text{constant}_k, \tag{2.3}$$

where $k = 1, \ldots, s$, for some constants α_{jk} and an integer s.

2.3 A mathematical model for protein gradients

In order to write the differential equations for the mathematical model that will translate a genetic network into a gradient of protein concentrations, we need to have the kinetic diagrams (2.1) that represent the interactions of the system, that will afterwards return the differential equations according to the law of mass action (2.2) and according to the conservations of the system (2.3). When these

equations are solved, with a numerically method in the large majority of the cases, the steady-state solutions will describe the experimental steady-state of protein concentration.

We start by defining a model for protein production from mRNA, and mRNA production from DNA. We consider that transcription factors may bind to a binding site associated to a given gene: if this transcription factor happens to be an activator, the gene transcription into mRNA is promoted; if this transcription factor is a repressor, the gene transcription is inhibited. The complex binding site plus gene will be called an *operon* - see figure 2.2 [27].

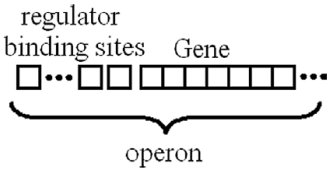


Figure 2.2: Jacob and Monod operon model where transcriptional regulators bind to the *regulator binging sites*, repressing or activating the translation of the *gene* into mRNA. Image taken from [27].

If we call G, M and P the gene, mRNA and protein, respectively, C_1 and C_2 catalysts that are not consumed in the reactions, and k_1, k_2, d_1 and d_2 the reaction rates, the simplest model of protein production from a DNA template without any regulation mechanism is:

$$\begin{aligned} C_1 + G &\xrightarrow{k_1} C_1 + G + M \\ C_2 + M &\xrightarrow{k_2} C_2 + M + P \\ M &\xrightarrow{d_1} \\ P &\xrightarrow{d_2} \end{aligned} \tag{2.4}$$

Furthermore, if we only consider protein degradation - that is, mRNA and catalysts which concentrations are constant - the above kinetic diagram can be simplified to:

$$\begin{aligned} G &\xrightarrow{k} G + P \\ P &\xrightarrow{d} \end{aligned} \tag{2.5}$$

We now consider the case where a transcription activator binds to an operon site, following the initiation of transcription and translation - the case of positive regulation. If we represent the activator's concentration by A, the concentration of the gene to be transcribed by G and the binded DNA-activator complex by G_A, the mechanism for protein production is:

$$\begin{aligned} A + G &\underset{k_2}{\overset{k_1}{\rightleftharpoons}} G_A \\ G_A &\xrightarrow{a} G_A + P \\ P &\xrightarrow{d} \\ A &\xrightarrow{d_A} \end{aligned} \tag{2.6}$$

where the last reaction may be omitted if we do not consider that the activator degrades.

We now consider the special case where the protein favors its own production, that is, a case of self-activation. In a similar manner, we say that G_P will be the DNA-activator complex and, for this case, the mechanism is:

$$
\begin{aligned}
&P + G \underset{k_2}{\overset{k_1}{\rightleftharpoons}} G_P \\
&G_P \xrightarrow{a} G_P + P \\
&P \xrightarrow{d}
\end{aligned}
\tag{2.7}
$$

The case for negative regulation will be analogous, but instead of considering an activator A, we consider a repressor R, and instead of considering the complex G_A, we consider G_R:

$$
\begin{aligned}
&R + G \underset{k_4}{\overset{k_3}{\rightleftharpoons}} G_R \\
&G \xrightarrow{a} G + P \\
&P \xrightarrow{d} \\
&R \xrightarrow{d_R}
\end{aligned}
\tag{2.8}
$$

Once again, just like in the positive regulation case, the last reaction is omitted if the repressor concentration is constant and not omitted if there is degradation. Analogously to the positive regulation case, we also consider the situation where the protein itself inhibits its own production - the case of *self-repression*. If then, the considered mechanism shall be [27]:

$$
\begin{aligned}
&P + G \underset{k_4}{\overset{k_3}{\rightleftharpoons}} G_P \\
&G \xrightarrow{a} G + P \\
&P \xrightarrow{d}
\end{aligned}
\tag{2.9}
$$

2.4 Reaction only model

In [22], an analytical model based on the mass action law (local reaction type) is shown to be sufficient to reproduce gap and maternal proteins regulatory network and their expression. [23] is also able to reproduce maternal, gap and pair-rule expression using a thermodynamic description for the genes regulation by transcription factors and, assuming independent contribution from multiple enhancers, it is able to calibrate experimental data and predict unknown parts of the regulatory network. The data prediction for *Eve*'s pattern is very accurate: the underlying regulatory network is show in figure 2.3 and the result is reproduced in figure 2.4.

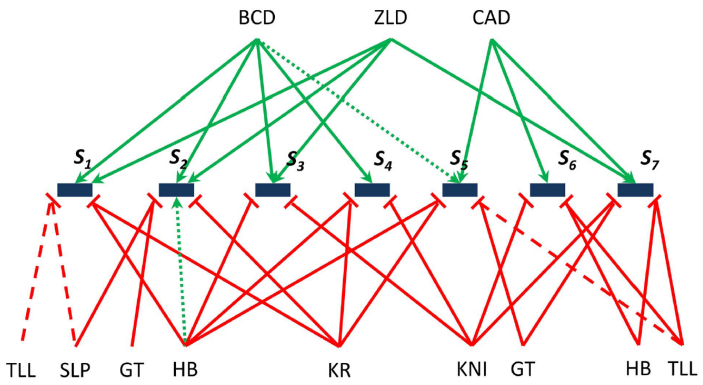


Figure 2.3: Networks of [23] showing regulatory influences of TFs on individual stripes of *Eve*. Red edges denote repressive and green edges denote activating role of the corresponding TF. Solid edges denote predicted influences in that are already known in the literature. Edges with large dashes denote predicted influences that were not reported in the literature before, while edges with small dashes denote predicted influences already known in literature but missed by the model (false negatives). Taken from figure 7 A of [23].

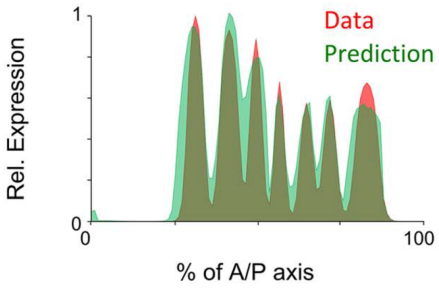


Figure 2.4: *Eve* model prediction at 14A8 by [23].

.

Nevertheless, the assumption that each stripe border is independently regulated by a given group of maternal and gap proteins [23] does not have a clear biological ground. It is possible that this is an *ad hoc* theory to explain the sudden increase of spatial frequency that we observe from gap to pair-rule proteins. Therefore, we keep the theory that *Eve* is the result of a set of local activations and repressions, but we consider that these mechanisms occur through all embryo length, which seems more biological plausible than having a specific set of controlling proteins for each stripe. Still, if we look at figure 1.4, we perceive some regularities between the gap genes final pattern and *Eve* repressed regions: *Kr* has a maximum between stripes 3 and 4, *Kni* has a maximum between stripes 5 and 6, *Gt* has a local maximum concentration between stripes 1 and 2, and then between stripes 6 and 7 and *Hb* increases its concentrations after stripe 7. This means that, although we abandon the hypothesis of different groups of promoters and repressors, the assumption of global activation or repression by the maternal products and a localized repression by the gap genes seems promising.

2.4.1 Implementation

Following the mathematical model for protein production in 2.3, we consider that each pair-rule gene is positively regulated - see mechanism (2.6) - by a set of n gap or maternal genes, $A_1(x), \ldots, A_i(x), \ldots, A_n(x)$, that have too a distribution over the embryo length x; moreover, we consider that the gene

is also negatively regulated - see mechanism (2.8) - by a set of m gap or maternal genes, $R_1, \ldots, R_m$. The set of activators can be replaced by an effective activator, $A(x) = A_1(x) + \cdots + A_i(x) + \cdots + A_n(x)$, and the set of activators can be replaced by an effective repressor, $R(x) = R_1(x) + \cdots + R_i(x) + \cdots + R_n(x)$. As *Eve* and *Ftz* proteins disapear later in development, we consider their degradation; on the contrary, since gap and maternal proteins are approximately constant in time, we do not consider the degradation neither of the effective activator nor of the effective repressor. This way, our global mechanism for the production of a pair-rule protein is going to be:

$$\begin{aligned} &G + A \underset{l_{-1}}{\overset{l_1}{\rightleftharpoons}} G_A \\ &G_A \xrightarrow{k_p} G_A + P \\ &G + R \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} G_R \\ &P \xrightarrow{\delta} \end{aligned} \tag{2.10}$$

According to the law of mass action (2.2), the time evolution equations are going to be:

$$\begin{aligned} A'(t) &= l_{-1}G_A(t) - l_1 A(t)G(t) \\ R'(t) &= k_{-1}G_R(t) - k_1 G(t)R(t) \\ P'(t) &= k_p G_A(t) - \delta P(t) \\ G'_A(t) &= -l_{-1}G'_A(t) + l_1 A(t)G(t) \\ G'_R(t) &= -k_{-1}G_R(t) + k_1 G(t)R(t) \\ G'(t) &= k_{-1}G_R(t) + l_{-1}G_A(t) - A(t)G(t)l_1 - G(t)R(t)k_1, \end{aligned} \tag{2.11}$$

with the conservation equations, according to (2.3):

$$\begin{aligned} A(t) - G(t) + R(t) &= A_0 - G_0 + R_0 \\ -A(t) + G(t) + G_R(t) &= -A_0 + G_0 \\ A(t) + G_A(t) &= A_0, \end{aligned} \tag{2.12}$$

where $R_0 = R(t=0)$, $A_0 = A(t=0)$ and $G_0 = G(t=0)$. We also consider that the transcription starts at $t = 0$, so that $G_R(t=0) = G_A(t=0) = P(t=0) = 0$. For the positive and negative regulations, we should have $l_1 \gg l_{-1}$ and $k_1 \gg k_{-1}$. We start by considering the evolution in time only, that is, the activator, the repressor and the gene initial concentration have a uniform distribution is space. The numerical solution of the equations for $A(t)$, $R(t)$ and $P(t)$ until steady-state ($t \approx 50$) is reached is depicted in figure 2.5.

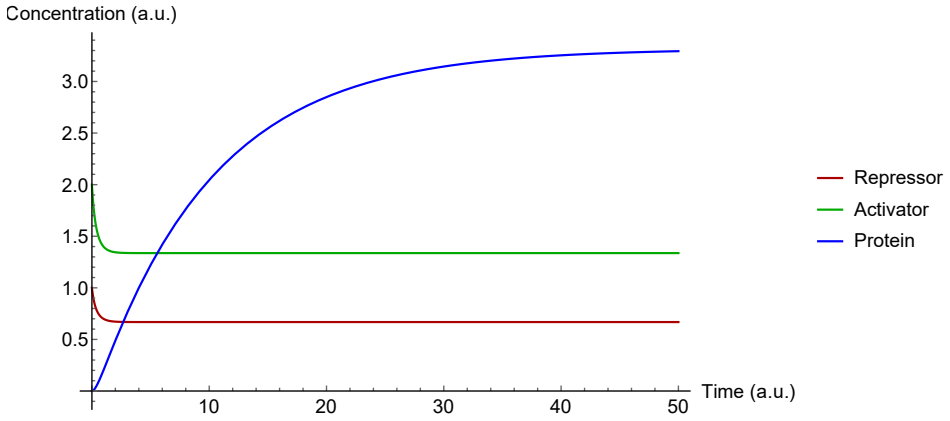


Figure 2.5: Evolution of the concentrations of the activator, repressor and protein, for $R_0 = 2.0$, $A_0 = 1.0$ and $G_0 = 1.0$, according to (2.11) and (2.12).

We then investigate how a pattern with a spatial distribution is altered by the action of an activator and a repressor. We propose an activator and a repressor with the initial spatial distribution of $A(x,0) = 0.5 + 0.1\exp -\frac{(x-0.8)^2}{0.001}$ and $R(x,0) = 0.5 + 0.3\exp -\frac{(x-0.5)^2}{0.001}$ (respectively), which are depicted in figure 2.6.

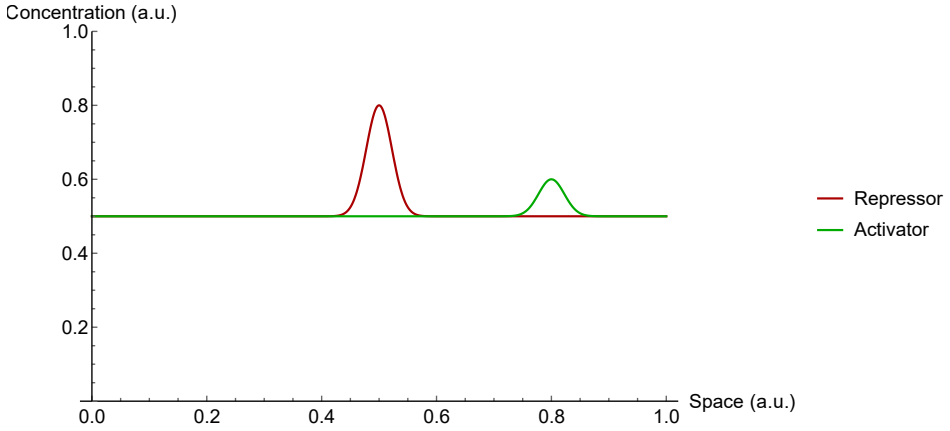


Figure 2.6: Proposal of an initial spatial distribution of the activator and the repressor. Code at *model_activators_repressors.nb*.

The steady-state distribution in space of the protein, at $t = 50$, can be seen in figure 2.7.

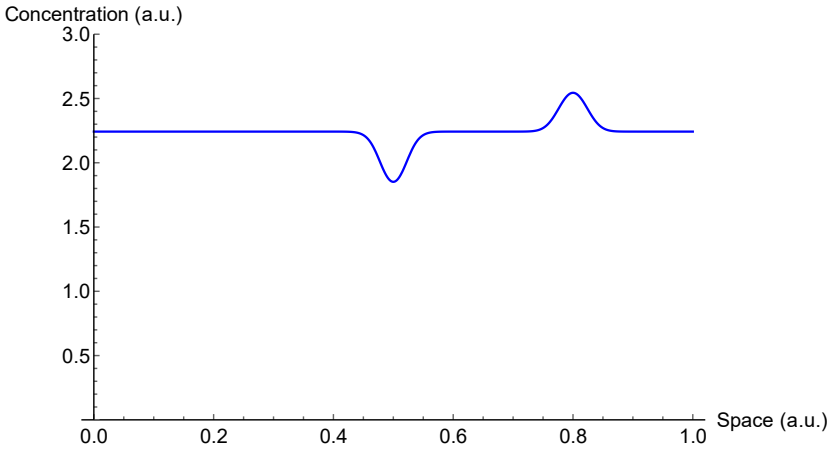


Figure 2.7: Solution of the steady state ($t = 50$) distribution of the protein for $G_0 = 1.0$, in the presence of an activator and a repressor with the initial distribution seen in figure 2.6, according to (2.11) and (2.12). Code at *model_activators_repressors.nb*.

The results are what we intuitively predict: the protein has an increasing of concentration at the position of the activator, and a decreasing of concentration at the position of the repressor.

2.4.2 Results

We use this model to fit the normalized *Eve* pattern at cycle 14A8, with the maternal proteins *Bcd* and *Cad*, plus the gap genes *Kr*, *Kni*, *Gt*, *Hb* and *Tll* as possible activators or repressors, with their profiles at cycle 14A1. For this fit, the rate constants of activation and repression will be proportional to a random number between 0 and 10. We started by making 500 combinations of this kind, subtracted the intercept, normalized the result and chose the combination which gave the best χ^2. If, for a given protein, its repressing rate was at least twice its activation rate at that combination, we concluded it was a repressor and, in the next set of random rate constants, that protein had its activation rate set at 0 (and vice-versa for an activator); if not, it would have a random rate constant between 0 and 10 once more. Each time we set a protein as an activator or repressor, we iterated the procedure twice the times we had before, and we stopped iterating when all proteins were fixed as either repressors or activators. The results are shown in figure 2.8.

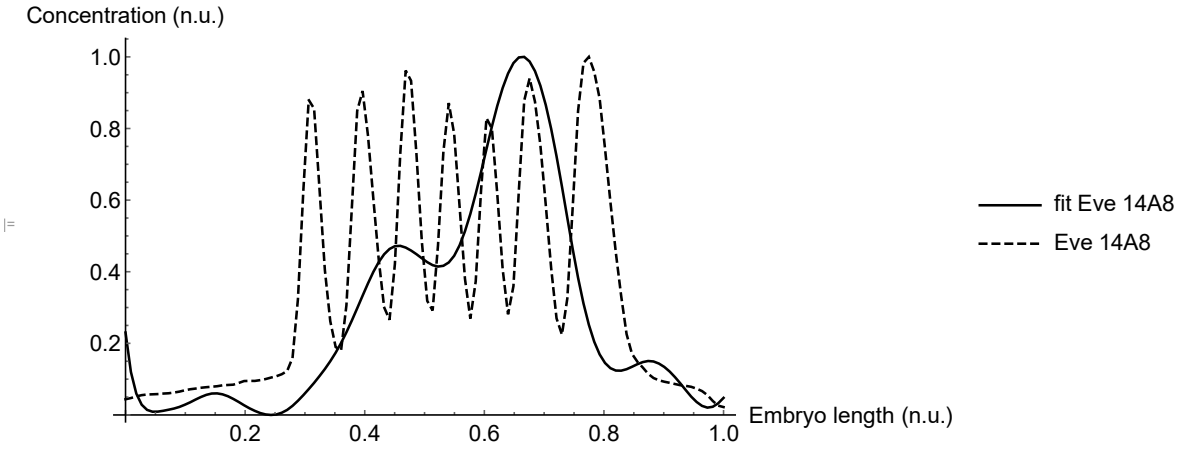


Figure 2.8: The best fit was obtained with $\chi^2 = 18.2$ after 8000 loops using 112 points, with the rate constants of table 2.1. Code at *model_activators_repressors_14A1_v2.nb*.

	Bcd	Cad	Kr	Kni	Gt	Hb	Tll
Activation		2.6		4.4		3.2	
Repression	8.9		1.2		6.8		2.1

Table 2.1: Activating and repressing proteins for the best fit, depicted in figure 2.8, and their respective rate constants ratios.

This way, we conclude that, although we perceive some regularities between the gap and the maternal proteins pattern, as well as correspondences between possible regions for activation and repression of the pair-rule stripes, these proteins' profiles are not sufficient to produce such a high frequency spatial pattern.

3

Reaction-diffusion model

Contents

3.1 Reaction-diffusion mechanisms

Alan Turing proposed in his seminal work that morphogenesis, which is the ensemble of processes that determines form, shape and patterns in organisms, is the result of a dynamical system that considers not only reaction but also diffusion [32].

We begin with an homogeneous system with the presence of irregularities, including statistical fluctuations in the number of molecules undergoing those various reactions, which are going to break this homogeneity in the presence of the appropriate kind of instability. This instability is going to trigger reaction and diffusion mechanisms, that may result in a steady-state patterns. Without the presence of these triggering instabilities no pattern is formed. Moreover, the pattern that is formed in the steady state may depend on form of the irregularities that prompted it. In what concerns diffusion, this *reaction-diffusion system* typically produces stable patterns when the two species diffuse at very different rates, nonetheless, systems with species diffusing at the same rate have been found too [33]. Here, we should note that the innovative aspect in this model is that diffusion is presented with an essential role in pattern formation, whereas it is typically thought to have a homogenizing effect as, for example, the dispersion of as ink droplet in water. A scheme of this process is presented in figure 3.1.

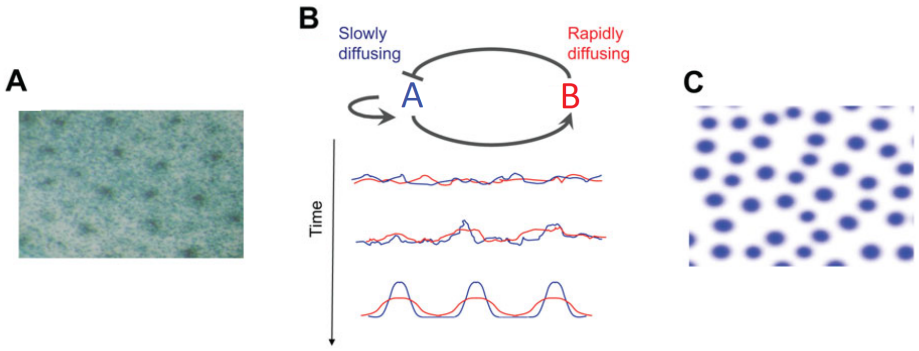


Figure 3.1: Example and illustration of the emergence of a Turing pattern. A: The system is homogeneous at first but there are some irregularities. B: Two species of this system are going to react with each other and diffuse as well, forming a spatial profile. C: The 2D distribution of the species and the 1D Turing pattern result in the formation of spots. Scheme adapted from [34].

3.1.1 Model for pattern formation

We shall then consider a system of n species, each jth species A_j with a diffusion coefficient D_j: the evolution of a species concentration is a result of a source term due to the reactions with all the other substances plus a term concerning the diffusion, according to the Fick's equation:

$$\frac{dA_j}{dt} = f_j(A_1, \ldots, A_n) + D_j \Delta A_j, \tag{3.1}$$

where $\Delta = \left(\frac{\partial^2}{\partial x_1^2} + \ldots + \frac{\partial^2}{\partial x_k^2}\right)$ is the k-dimensional Laplace operator. From now on, we will consider a 1-dimensional system, *i.e.* $k = 1$. If the jth substance diffuses (that is, $D_j > 0$) we say that this substance is a *morphogene*, *form producer* or *evocator*. If $D_j = 0$, we identify this species as a non-diffusing substance [30].

3.2 The Brusselator model

The Brusselator is a simple model able to produce Turing patterns with two species. It mimics an autocatalytic process and has the following kinetic mechanism:

$$\begin{aligned} &\mathrm{A} \xrightarrow{\mathrm{k_1}} \mathrm{X} \\ &\mathrm{B} + \mathrm{X} \xrightarrow{\mathrm{k_2}} \mathrm{Y} + \mathrm{D} \\ &2\,\mathrm{X} + \mathrm{Y} \xrightarrow{\mathrm{k_3}} 3\,\mathrm{X} \\ &\mathrm{X} \xrightarrow{\mathrm{k_4}} \mathrm{E}, \end{aligned} \tag{3.2}$$

where X is the autocatalytic species. If we apply the law of mass action (2.2), we obtain the differential equations which will determine the reaction part of the reaction-diffusion system:

$$\begin{aligned} \frac{dX}{dt} &= k_1 A - k_2 B X + k_3 X^2 Y - k_4 X \\ \frac{dY}{dt} &= k_2 B X - k_3 X^2 Y \\ \frac{dA}{dt} &= -k_1 A \\ \frac{dB}{dt} &= -k_2 B X \\ \frac{dE}{dt} &= k_4 X \\ \frac{dD}{dt} &= k_2 B X. \end{aligned} \tag{3.3}$$

According to (2.3), this system obeys to the conservation laws:

$$\begin{aligned} &B(t) + D(t) = B(0) + D(0) \\ &X(t) + Y(t) + A(t) + E(t) = X(0) + Y(0) + A(0) + E(0). \end{aligned} \tag{3.4}$$

If we now assume that A and B are constants, we obtain the following one-dimensional reaction-diffusion system for species X and Y, taking the first two equations from (3.3) as the reaction terms:

$$\boxed{\begin{aligned} \frac{\partial X(x,t)}{\partial t} &= k_1 A - k_2 B X(x,t) + k_3 X^2(x,t) Y - k_4 X(x,t) + D_X \frac{\partial^2 X(x,t)}{\partial x^2} \\ \frac{\partial Y(x,t)}{\partial t} &= k_2 B X(x,t) - k_3 X(x,t)^2 Y(x,t) + D_Y \frac{\partial^2 Y(x,t)}{\partial x^2}, \end{aligned}} \tag{3.5}$$

where D_X and D_Y are the diffusion coefficients for species X and Y respectively.

The local component of the vector field associated to the equation (3.5) has one fixed point at $(X_0, Y_0) = (A\frac{k_1}{k_4}, \frac{B}{A}\frac{k_2 k_4}{k_1 k_3})$ and a supercritical Hopf bifurcation for $B = B_{HB} = \frac{k_4}{k_3} + A^2 \frac{k_1^2 k_3}{k_2 k_4^2}$, so that if $B < B_{HB}$ the fixed point is a stable focus, and if $B > B_{HB}$ the fixed point is an unstable focus.

For the numerical integration of the system, we use Euler's method for time evolution and finite differences for space evolution - for the latter, we consider Neumann boundary conditions (zero flux). We used $dt = 0.001$ for Euler's method step and $dx = \sqrt{\frac{max(D_X, D_Y)}{\gamma}} dt$ for the finite differences, with $\gamma = \frac{1}{6}$, since this relationship is reported to lead to optimal convergence to the solution of the system [35]. In figure 3.2, we present the bifurcation diagram of the solutions of the Brusselator model (3.5),

for the parameter values $A = 2$, $k_1 = k_2 = k_3 = k_4 = 1$ and $D_2 = 1$: Turing patterns, at the dark grey region, only appear when we have $D_1 > D_2$ and arise on both sides of the Hopf bifurcation.

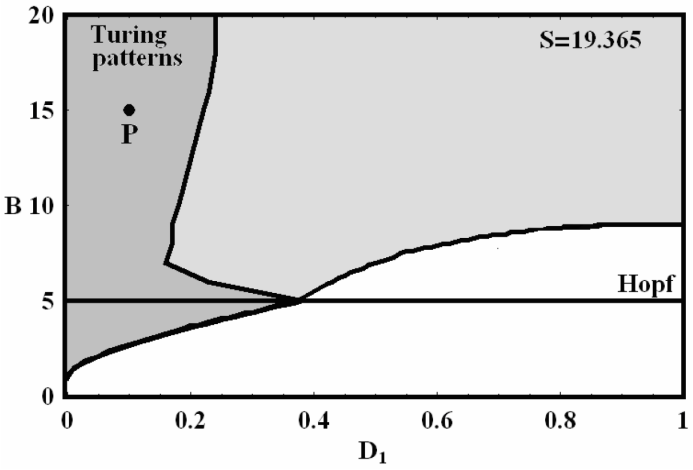


Figure 3.2: Bifurcation diagram of the solutions of the Brusselator model (3.5), for the parameter values $A = 2$, $k_1 = k_2 = k_3 = k_4 = 1$ and $D_2 = 1$, in a one-dimensional domain of length $S = 19.365$. Taken from figure 2 of [30].

In figure 3.3, we reproduce the Turing pattern obtained at P of diagram 3.2, that is, with the parameters

$$\begin{aligned} A &= 2.0 \\ B &= 15.0 \\ D_X &= 0.1 \\ D_Y &= 1.0 \\ k_1 &= k_2 = k_3 = k_4 = 1, \end{aligned} \tag{3.6}$$

and, for this arrangement of parameters, the local system (3.3) has one unstable fixed point and one limit cycle in the phase space. We chose to simulate the system with $M = 250$ lattice sites, thus the length of the spatial domain is $M\sqrt{\frac{\Delta t}{\gamma} max(D_X, D_Y)} \approx 19.52$; the simulation was run until the concentrations of both species in all lattice sites were less that 0.001% different that its previous value (time), and this was our criterion to detect the steady-state while taking into account small numerical deviations. Moreover, for the instabilities that are going to trigger the reaction-diffusion mechanisms we chose an initial profile with the values at the fixed point plus a random profile $\delta(x)$ with values between 0.0 and 0.5, $X(0, x) = \frac{k_1 A}{k_4} + \delta(x)$ and $Y(0, x) = \frac{k2\, k4\, B}{A\, k1\, k3} + \delta(x)$[30].

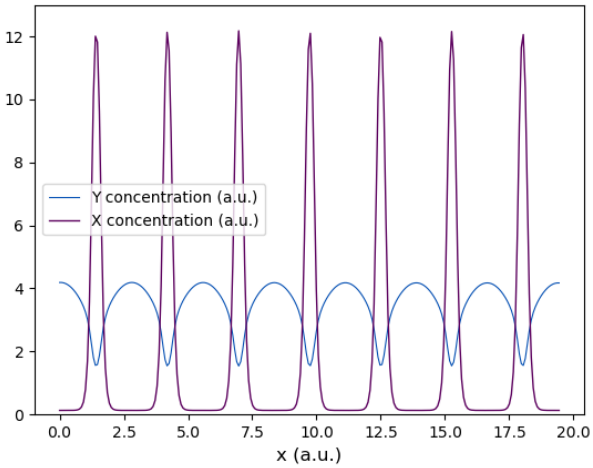


Figure 3.3: Turing pattern of the Brusselator reaction-diffusion system (3.5) with parameters (3.6). Code at *brusselator_cycle14_copy2.py*.

3.3 Reaction-diffusion model

3.3.1 Development of the initial pattern

Since it was not possible to form a pair-rule's pattern via local activation or repression by the gap and maternal genes, we now propose that these early stage proteins are only responsible for the setting of the initial pattern, which then develops stripes via a Turing mechanism. This hypothesis has already been proposed in [36] and [37]. Moreover, [6] also suggests that these two pair-rule genes are subject to positive auto-regulation control, and [38] and [39] suggest that the stripe-like activation of genes is a result of an autocatalytic feedback, which is experimentally supported as we mentioned in section 1.2.2 of chapter 1.

For the setting of the initial pattern, and following the method described in 2.4.2, we fitted *Eve* pattern at cycle 14A1: the result is shown in figure 3.4 and the reaction constants in table 3.1.

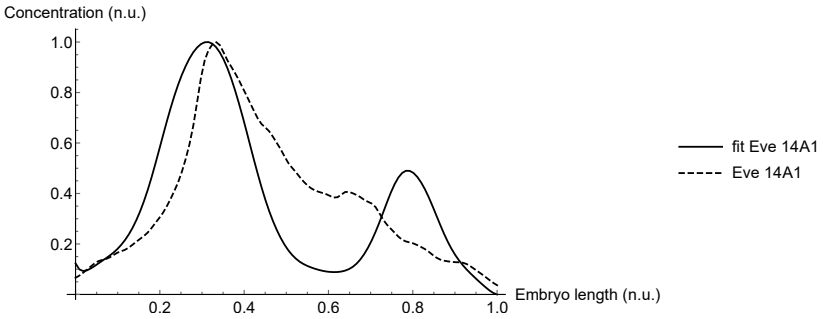


Figure 3.4: The best fit was obtained with $\chi^2 = 20.7$ after 8000 loops, using 112 points, with the rate constants of table 3.1. The experimental profile is normalized. Code at *model_activators_repressors_14A1_RD.nb*.

	Bcd	Cad	Kr	Kni	Gt	Hb	Tll
Activation					0.9	0.3	
Repression	9.4	4.3	0.4	0.8			6.4

Table 3.1: Activating and repressing proteins for the best fit, depicted in figure 3.4, and their respective rate constant ratios.

We have obtained the anterior protein *Bcd* and the posterior proteins *Cad* and *Tll* as major repressors, whereas the other gap genes have a residual influence on this pattern formation, and we also observe an odd concavity at $x \approx 0.6$, caused by *Tll* repression. The fact that *Bcd* and the majoriity of gap genes are repressors meets very well the experiments with mutations described in chapter 1.2, where the mutations of these caused and enlargement of *Eve* stripes. The exceptions are the gap genes *Gt* and *Hb*, which are here identified as activators instead of repressors. We got a very good fit for the anterior part of the profile, until $x \approx 0.3$, where the concentration has its maximum, as the intersept and the second derivatives (positive at first and negative at last) very closely match the experimental data. To further analyze the strange concavity at $x \approx 0.6$, we check how the pattern of *Eve* at 14A1 would be, in general terms, with *Bcd* repression only - figure 3.5 - and then with the repressors *Bcd* and *Cad* - figure 3.6 .

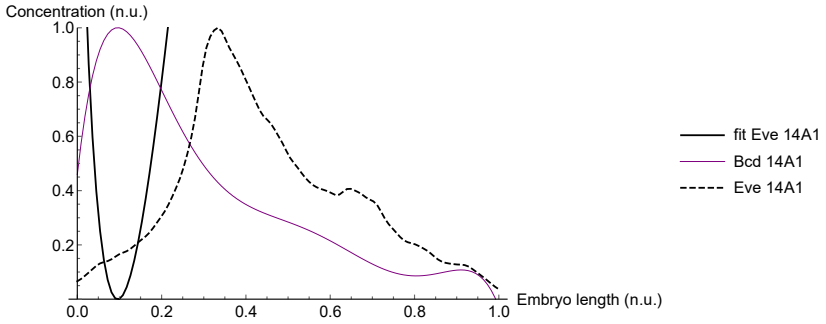


Figure 3.5: Test profile ("fit Eve 14A1") if we had a constant activator and *Bcd* as a repressor. Code at *model_activators_repressors_14A1_RD.nb*.

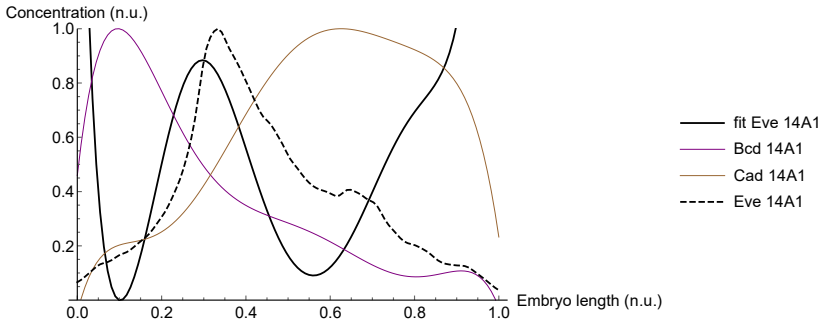


Figure 3.6: Test profile ("fit Eve 14A1") if we had a constant activator plus *Bcd* and *Cad* as repressors, with the same reaction constants. Code at *model_activators_repressors_14A1_RD.nb*.

From fit 3.5, we see that *Bcd* is (almost) sufficient to produce the anterior part of the profile alone, except for the very anterior end $x < 0.1$. If we add *Cad* as a repressor, see figure 3.6, our profile meets the experimental one very well until $x \approx 0.6$. In fact, this test profile diverges immensly at the edges ($x < 0.1$ and $x > 0.9$), because of a sudden change in the derivative of the repressors *Bcd* and

Cad, where their concentrations start decreasily abruptly. This is a systematic error we will have in these regions, and it is due to a shadow present in all photographs from which we extracted the data - see figure 3.7.

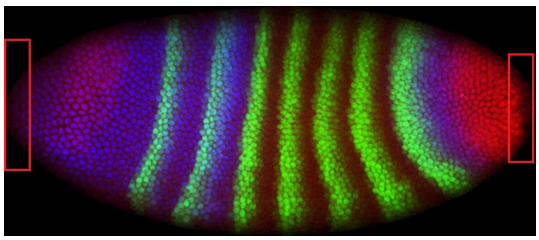


Figure 3.7: Example of a FlyEx database file - notice the shadow at the embryo's border.

Thereafter, we fitted *Ftz* pattern at cycle 14A1: the result is shown in figure 3.8 and the reaction constants in table 3.2.

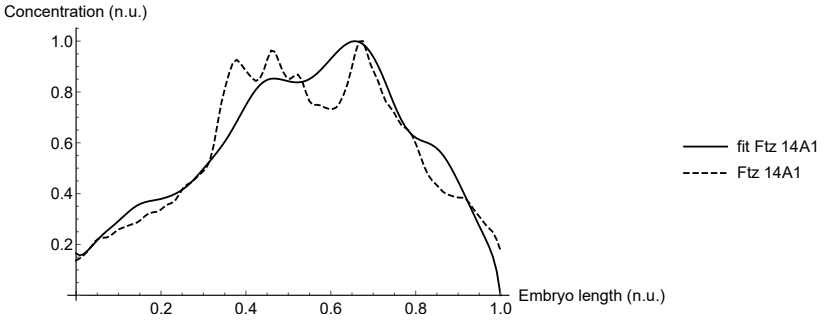


Figure 3.8: The best fit was obtained with $\chi^2 = 1.7$ after 16000 loops using 112 points, with the rate constants of table 3.2.The experimental profile is normalized. Code at *model_activators_repressors_14A1_RD.nb*

	Bcd	Cad	Kr	Kni	Gt	Hb	Tll
Activation		9.1	0.5	3.5		6.7	
Repression	9.2				6.0		5.9

Table 3.2: Activating and repressing proteins for the best fit, depicted in figure 3.8, and their respective rate constant ratios.

In what concerns *Ftz*, our fit very closely matches the experimental data, and we once again obtained *Bcd* as a major repressor, and *Cad* is now an activator instead. Moreover, the gap genes *Kni*, *Gt*, *Hb* and *Tll* no longer have a residual influence on the profile initiation. Nonetheless, the fit does not reproduce the small irregularities in central part: this suggests that, even at this early stage, the proteins *Eve* and *Ftz* have already started interacting.

In summary, these eary stage fits improved very much when compared to the last stage fit at figure 2.8, which means our reaction-diffusion hypothesis may have a closer match to reality.

3.3.2 Development of stripes

As the initial pattern of *Eve* and *Ftz* are already established, we propose that these two proteins start interacting with each other, using the Brusselator (3.5) as a template for this reaction, where the two species that feed the system, A and B, are going to be a combination of maternal and gap genes.

Since this combination alone is going to set the initial pattern for *Eve* and *Ftz*, we are the going to use $A(x) = Eve14A1$ and $B(x) = Ftz14A1$. We tested all the combinations of 0.01, 0.1 and 1.0 for the diffusion coefficients of species X (D_X) and Y (D_Y), and we obtained three different patterns, which are depicted in figures 3.9, 3.10 and 3.11.

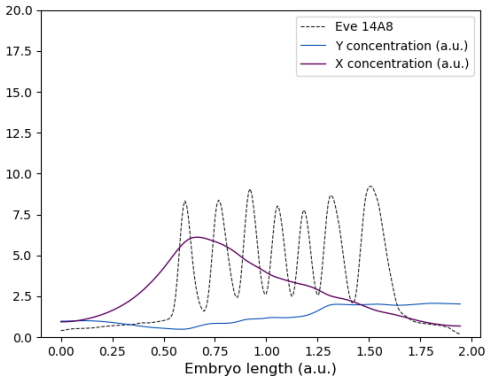


Figure 3.9: Steady-state pattern of (3.5) using $A(x) = Eve14A1$ and $B(x) = Ftz14A1$. This pattern was obtained for $D_Y = 0.01$, $(D_X, D_Y) = (0.1, 0.1)$,$(1.0, 0.1)$,$(1.0, 1.0)$.

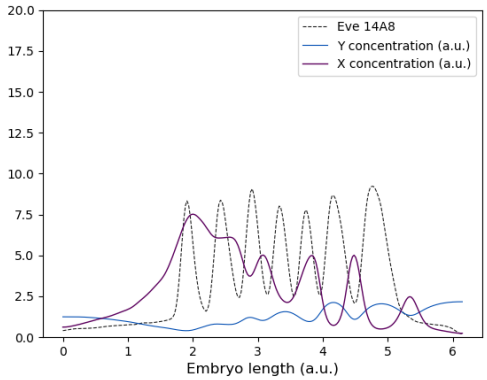


Figure 3.10: Steady-state pattern of (3.5) using $A(x) = Eve14A1$ and $B(x) = Ftz14A1$. This pattern was obtained for (D_X, D_Y)=$(0.01, 0.1)$,$(0.1, 1.0)$.

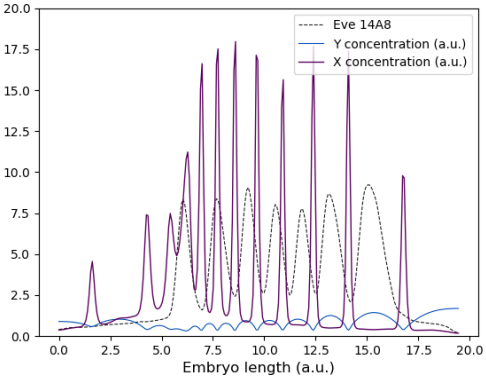


Figure 3.11: Steady-state pattern of (3.5) using $A(x) = Eve14A1$ and $B(x) = Ftz14A1$. This pattern was obtained for (D_X, D_Y)=$(0.01, 1.0)$. Code at *brusselator_XY_eve14A1_ftz14A1.py*.

This way, we conclude that we only obtain a Turing pattern (3.10 and 3.11) when D_Y is 10 or 100 bigger than D_X, which was expected since in the majority of Turing patterns one diffusion coefficient is way larger than the other. More specifically, if we fix $D_Y = 1.0$, there will be Turing patterns for (at least) $D_X \in [0.01, 0.1]$, where $D_X = 0.01$ gives 6 stripes (figure 3.10) and $D_X = 0.1$ gives 12 (figure 3.11), and the number of stripes for other intermediary D_X can be checked in table 3.3.

$D_X(\times 10^{-2})$	1	2	3	4	5	6	7	8	9	10
Nr. stripes	12	10	8	7	7	7	6	6	6	6

Table 3.3: Number of stripes of the simulated pattern as a function of D_X, for $D_Y = 1.0$.

Therefore, we find a 7 stripe pattern for $D_X \in [0.04D_Y, 0.06D_Y]$, and the pattern for these is depicted in figure 3.13. In order to explore if this is a strict or a reasonable range for the diffusion coefficient, further experiments to measure this constant are necessary. In figure 3.12 we recall *Eve* and *Ftz* steady-state pattern, and in figure 3.14 we show the same pattern as in 3.12 but with Y pattern at a more similar scale to X for comparison.

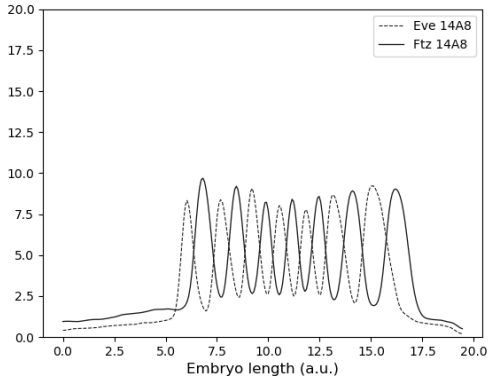


Figure 3.12: *Eve* and *Ftz* protein patterns at cycle 14A8. Code at *brussela-tor_XY_eve14A1_ftz14A1_plot.py*

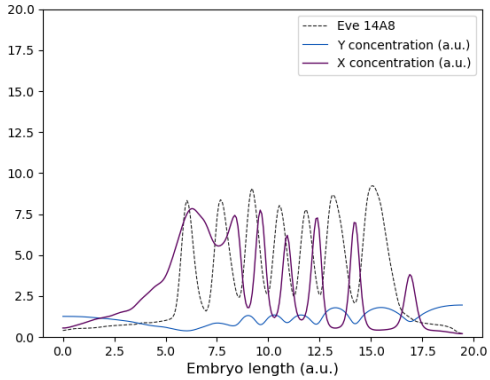


Figure 3.13: Steady-state pattern of (3.5) using $A(x)$=Eve14A1, $B(x)$=Ftz14A1 and (D_X, D_Y)=$(0.05, 1.00)$. Code at *brussela-tor_XY_eve14A1_ftz14A1.py*

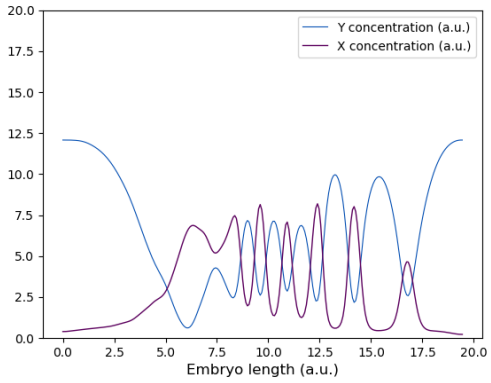


Figure 3.14: Same as figure 3.13, but with pattern of Y at a different scale for better comparison.

For the identification with the time evolution of figure 1.3, in figures 3.15 and 3.16 we show eight frames of the integration of the system for which the steady-state is depicted in figure 3.13, separated in equal time intervals. The first four stages match quite closely the experimental data: the maximum concentration increases from 1 to 2 and is constant from that time onwards, and the first

stripe projections appear in stage 2 in both experimental data and simulations. In the pattern which derives from the Brusselator model, these projections begin to elongate, forming stripes, like what happens in the *Eve* evolution. This way, we conclude that a model that considers reaction-diffusion, with auto-catalysis and interactions for the two species, suits the biological formation of *Eve* stripes.

We also supposed that the species Y with which X interacts and forms a Turing patterns would be the pair-rule complementary *Ftz*. Nonetheless, the Y concentration pattern at cycle 14A8 in figure 3.13 does not exactly match the steady-state pattern of *Ftz* (see figure 3.12): although both patterns have the same maximums and minimums (that is, both are complementary with *Eve* or X), their second derivatives have opposite signs in the striped region and, in the side regions without the pattern (for embryo length $\sim< 5.0$ and $\sim> 17.5$), Y concentration decreases towards the center, while *Ftz* increases towards the center. This means that the interaction term in the Brusselator model $\pm X^2Y$ does not describe the pair-rule interaction, and a correct identification of the dynamical system is required. Furthermore, this interaction is essential to form both species' stripes, and it was observed experimentally that *Eve* and *Ftz* can form their patterns without each other, which means that these two proteins may interact with other pair-rules.

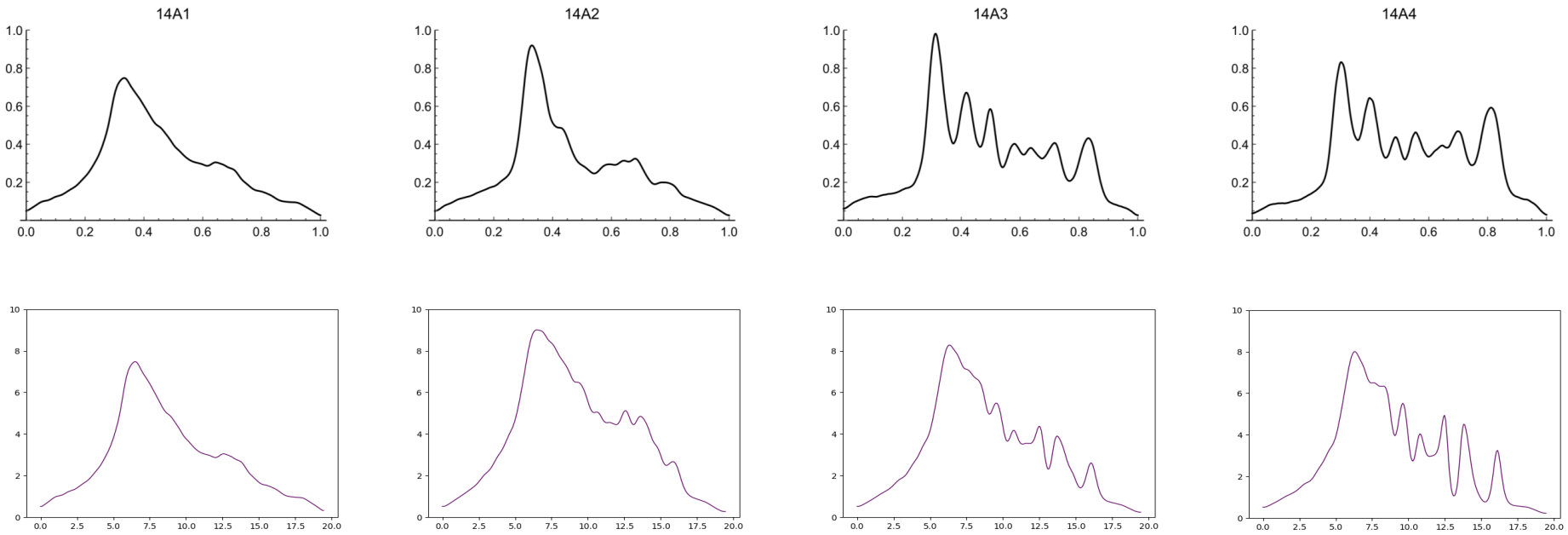


Figure 3.15: Normalized oncentration through normalized embryo length. *Eve* experimental patterns for cycles 14A1, 14A2, 14A3 and 14A4 (top figures) and comparison with four initial frames from the numerical intergation of the Brusselator model (3.5) (bottom figures), using $A(x)$=Eve14A1, $B(x)$=Ftz14A1 and $(D_X, D_Y) = (0.05, 1.00)$. Code at *brusselator_XY_eve14A1_ftz14A1_frames.py* and scheme at *frames_simulation_data.pptx*.

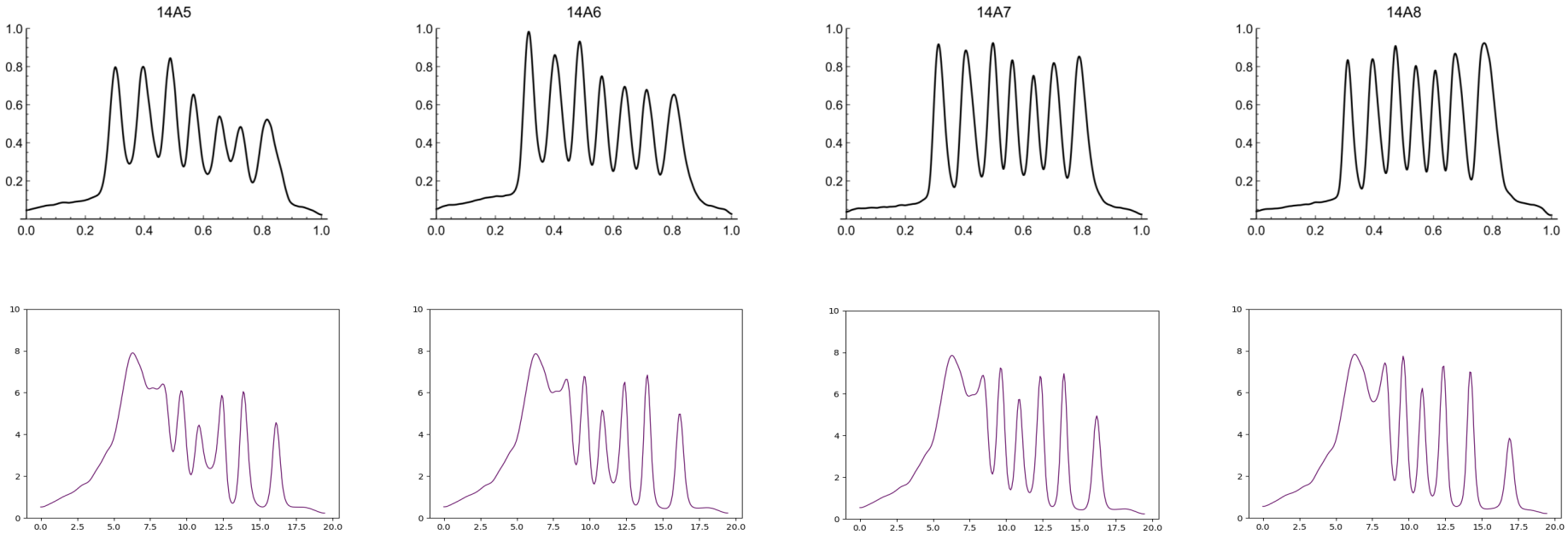


Figure 3.16: Normalized concentration through normalized embryo length. *Eve* experimental patterns for cycles 14A5, 14A6, 14A7 and 14A8 (top figures) and comparison with four final frames from the numerical intergation of the Brusselator model (3.5) (bottom figures), using $A(x)$=Eve14A1, $B(x)$=Ftz14A1 and $(D_X, D_Y) = (0.05, 1.00)$.

3.3.3 Segment-polarity stripes

We also propose a similar mechanism for the formation of the segment-polarity stripes. These are the last genes to be expressed in the *Drosophila* embryo, and their establishment occurs during the late cellular blastoderm stage (recall scheme 1.6), such that their pattern should be determined by the pair-rule proteins, and they have twice the number of stripes (fourteen) as these (seven). Studies have shown that one of these segment polarity, *en*, is activated by both *Eve* and *Ftz*, whereas another segment-polarity protein that develops at its side, *wg*, is repressed by the same pair-rule [7]. These two proteins are reported to react with each other and diffuse [40], what is also consistent with the Turing pattern formation mechanism. Moreover, in *Eve* and *Ftz* embryo mutants, *en* and *wg* have few and broader stripes than in wild type, which tells us that these pair-rule have an important role in these segment-polarity setting [7].

Analogously to what we did with the development of the pair-rule, we propose that these initialize the segment polarity pattern, which then evolves into stripes via a reaction-diffusion mechanism. With the same model for transcription and translation we used for the initiation of *Eve* in section 2.4.1, we present in figure 3.17 a protein (at green) which was activated by *Eve* and *Ftz*, in equal proportions, and another proteins (at red) which was repressed by the same pair-rule. Accordingly, these may mimic the initial pattern of *wg* and *en* respectively.

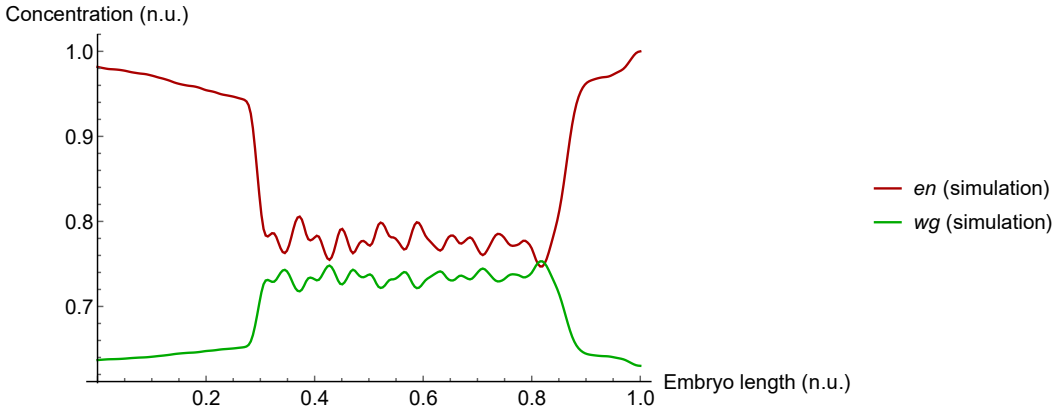


Figure 3.17: Possible initial pattern for *en* in red (repressed by *Eve* and *Ftz*) and for *wg* in green (activated by same proteins). Code at *en_wg_init.nb*.

For the reaction-diffusion part which is going to refine the stripes, we use the Brusselator as a template model once again plus $A(x) = wg$ and $B(x) = en$. We tested all the combinations of 0.01, 0.1 and 1.0 for the diffusion coefficients D_X and D_Y, and we obtained three different patterns, which are depicted in figures 3.18, 3.19 and 3.20.

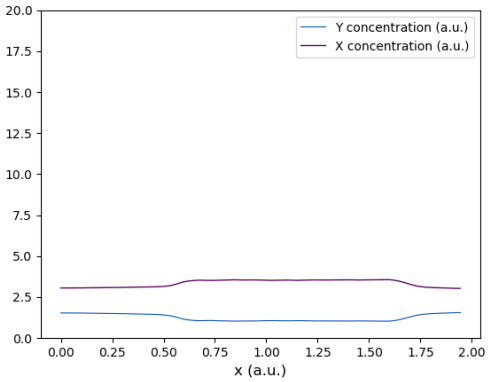


Figure 3.18: Steady-state pattern of (3.5) using $A(x) = wg$(simulation) and $B(x) = en$(simulation). This profile was obtained for $D_X = 1.0$, (D_X, D_Y)=$(0.1, 0.1)$, $(0.1, 0.01)$ and $(0.01, 0.01)$.

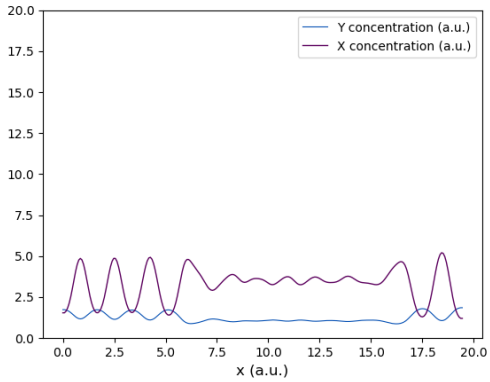


Figure 3.19: Steady-state pattern of (3.5) using $A(x) = wg$(simulation) and $B(x) = en$(simulation). This profile was obtained for $(D_X, D_Y) = (0.1, 1.0)$ and $(0.01, 0.1)$.

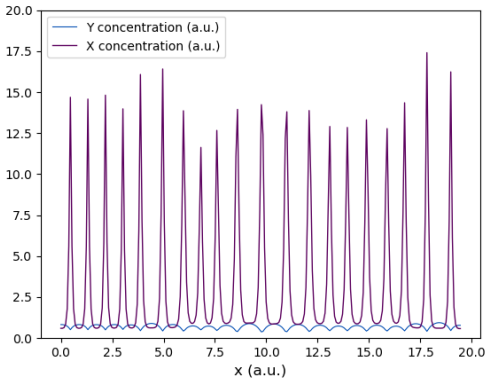


Figure 3.20: Steady-state pattern of (3.5) using $A(x) = wg$(simulation) and $B(x) = en$(simulation). This profile was obtained for $(D_X, D_Y) = (0.01, 1.0)$. Code at *brusselator_cycle14_original.py*

As it happened before with the pair-rule, we obtained a Turing pattern when D_Y is 10 or 100 bigger than D_X. For this reason, we looked for Turing patterns once more fixing $D_Y = 1.0$ and searched in the range $D_X \in [0.01, 0.1]$, and we could only find 14 stripes for $D_X = 0.03$, for which steady-state profile is depicted in figure 3.21. We once more suggest that this protein's diffusion coefficient should be measured in order to evaluate this result.

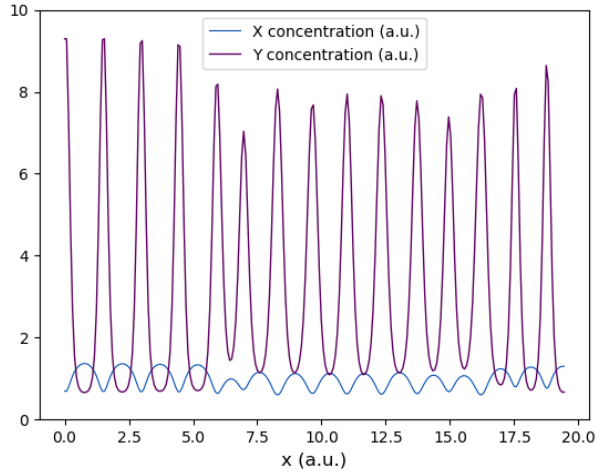


Figure 3.21: Steady-state pattern of (3.5) using $A(x) = wg$(simulation) and $B(x) = en$(simulation). This profile was obtained for $(D_X, D_Y) = (0.03, 1.0)$.

We should note that, in *en*- embryos, *wg* shows a broader ectopic transcription (vice-versa for *en* and *wg*- mutants), nonetheless, the same results are obtained for mutant embryos of *nkd* an *ptc*, which means that the refinement and maintenance processes does not happen only between these two, just like we proposed for the *Eve* and *Ftz* development [4].

4

Identification of the dynamical system

Contents

4.1 The SINDy algorithm

In order to understand a system behavior and regulation, accurately predict its outcomes and generalize its basic mechanisms for similar structures, it is essential to comprehend the system's fundamental dynamical relationships. In the present chapter, we aim to identify the system which regulates the pair-rule formation by taking advantage of an algorithm that gives a sparse identification of nonlinear dynamical systems - this is the SINDy algorithm, proposed by S. L. Brunton and colleagues [41]. In their work, the authors combine sparsity-promoting techniques and machine learning with nonlinear dynamical systems to discover governing equations from input data, where the only assumption made on the system is that its equations are sparse in the number of terms, which is reasonable to expect for the great majority of physical systems. The algorithm, which we now proceed to detail, provides candidate parsimonious models that balance accuracy with complexity, therefore avoiding overfitting.

4.1.1 Implementation

To construct this method we start by considering a dynamical system of the form

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t)), \tag{4.1}$$

where $\mathbf{x}(t) \in \mathbb{R}^n = [x_1(t)\, x_2(t) \ldots x_n(t)]^T$ is the state of the system at time t and $\mathbf{f}(\mathbf{x}(t))$ are the dynamic equations of the system. Thereafter, we collect a series of data at m time instances in order to construct the matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^T(t_1) \\ \mathbf{x}^T(t_2) \\ \vdots \\ \mathbf{x}^T(t_m) \end{bmatrix} = \begin{bmatrix} x_1(t_1) & x_2(t_1) & \ldots & x_n(t_1) \\ x_1(t_2) & x_2(t_2) & \ldots & x_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_m) & x_2(t_m) & \ldots & x_n(t_m) \end{bmatrix} \tag{4.2}$$

and calculate the numerical derivatives in the matrix

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{\mathbf{x}}^T(t_1) \\ \dot{\mathbf{x}}^T(t_2) \\ \vdots \\ \dot{\mathbf{x}}^T(t_m) \end{bmatrix} = \begin{bmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \ldots & \dot{x}_n(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \ldots & \dot{x}_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{x}_1(t_m) & \dot{x}_2(t_m) & \ldots & \dot{x}_n(t_m) \end{bmatrix}. \tag{4.3}$$

Since we aim to construct a nonlinear dynamical system for the model, we choose a set of p candidate functions, that consist in linear and nonlinear combinations of the data $\mathbf{x}(t)$, in which are the few terms that will give the dynamical regime of the system. If, for instance, the nonlinear candidate functions we choose are $x_1^2(t), \ldots, x_n^2(t)$ and $\cos(x_1(t)), \ldots, \cos(x_n(t))$, we then shall construct a $\Theta(\mathbf{X}) \in \mathbb{R}^{m\times p}$ with the candidate functions at the several m time instances:

$$\Theta(\mathbf{X}) = \begin{bmatrix} x_1^2(t_1) & \ldots & x_n^2(t_1) & \cos(x_1(t_1)) & \ldots & \cos(x_n(t_1)) \\ x_1^2(t_2) & \ldots & x_n^2(t_2) & \cos(x_1(t_2)) & \ldots & \cos(x_n(t_2)) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^2(t_m) & \ldots & x_n^2(t_m) & \cos(x_1(t_m)) & \ldots & \cos(x_n(t_m)) \end{bmatrix}. \tag{4.4}$$

This way, our candidate dynamical system, with all possible (within the chosen set) functions, is going to be given by $\dot{\mathbf{X}} = \Theta(\mathbf{X})$ but, since, only a few functions are going to be active in our system, we need to determine the sparse matrix of coefficients $\Xi \in \mathbb{R}^{p \times n}$, so that the real dynamical system is going to be a solution of:

$$\dot{\mathbf{X}} = \Theta(\mathbf{X})\Xi. \tag{4.5}$$

Alternatively, we can also compute the second derivative of $\mathbf{X}$ in (4.2), so that the input data is solution of the system [41]:

$$\ddot{\mathbf{X}} = \Theta(\mathbf{X})\Xi. \tag{4.6}$$

Hence, the key approach in this model is that it enables us to fit nonlinear functions through a linear fit.

4.1.1.A Determining the sparse matrix Ξ and sparsification parameter λ

We now proceed to explain how the Ξ matrix is determined. Firstly, we calculate Ξ with the method of least squares - in practice, we compute $\Xi = \Theta^{+}(\mathbf{X})\ddot{\mathbf{X}}$, where $\Theta^{+}(\mathbf{X})$ is the Moore-Penrose inverse, commonly refered as pseudoinverse. Then, we threshold all Ξ coefficients that are smaller then a chosen sparsification parameter λ. Once again, we obtain another least-squares solution for Ξ onto the remaining non-zero indices - this procedure is continued until the coefficients converge.

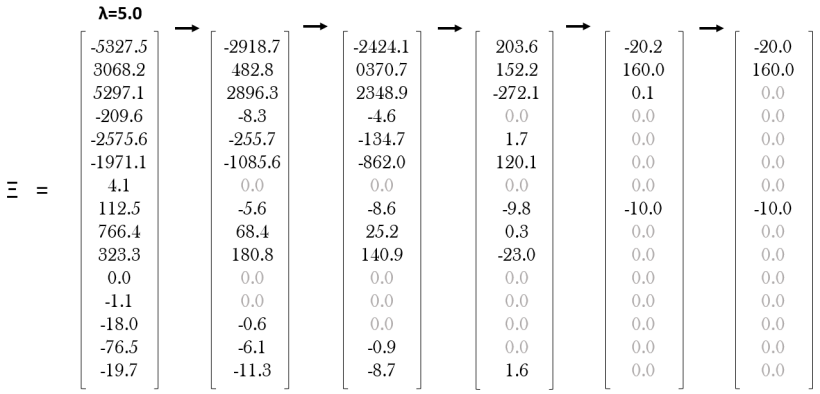


Figure 4.1: Example representing the sequential thresholding method for determining Ξ.

Moreover, the authors of [41] propose to determine the parameter λ as the one which represents a compromise in both accuracy and complexity (*i.e.*, number of terms in the equations). This way, λ should be at the Pareto front of these two quantities.

4.2 Identification of the Brusselator steady-state Turing pattern

The SINDy model is reported to be effiecient to determine the dynamical system of the damped oscillator, of the Lorenz system and of the Navier-Stokes equations. As benchmarking, we test if this algorithm is able to identify the equations for the steady-state Turing pattern of the Brusselator model, depicted in figure 3.3. Since this pattern is achieved at a steady-state, the solution verifies a

one-dimensional system according to (3.5):

$$\begin{aligned}\frac{d^2X(x)}{dx^2} &= -\frac{1}{D_X}\left(k_1A - k_2BX(x) + k_3X^2(x)Y - k_4X(x)\right)\\ \frac{\partial^2Y(x)}{\partial x^2} &= -\frac{1}{D_Y}\left(k_2BX(x) - k_3X(x)^2Y(x)\right).\end{aligned} \tag{4.7}$$

Choosing the same parameters as before (3.6), the system in these equations reads:

$$\begin{aligned}\frac{d^2X(x,t)}{dx^2} &= -20.0 + 160.0X(x) - 10.0X^2(x)Y(x)\\ \frac{d^2Y(x)}{dx^2} &= -15.0X(x) + X^2(x)Y(x).\end{aligned} \tag{4.8}$$

We now import the data of the steady-state pattern into the SINDy algorithm in order to see how it detects the linear and non-linear functions of this system. Firstly, we take the second derivatives for the altered SINDy model, so that we apply the proposal (4.6): taking into account the Neumann (zero flux) boundary conditions, we use central derivatives for the input of SINDy equation (4.6). We used as candidate functions combinations until 4^{th} order of X and Y. Running the algorithm for several values and orders of magnitude of the cut-off parameter λ, we found that we could only achieve a convergence to the solution for one species at a time.

In what concerns species X, if we compare the second derivatives computed with the SINDy output functions - see table 4.1 - and the ones computed directed from the data, we obtain a fit quality of $\chi^2 = 0.011$. Here we note that we cannot compare directly the data computed by the SINDy model with the original data because the evolution of the system depends on the initial profile, therefore it cannot be integrated. The values in table 4.1 were obtained for a range of sparsification parameters, which were $\lambda \in [1.2, 6.7]$.

$$\lambda \in [1.2, 6.7],\ \chi^2_X = 0.011$$

	X	X - SINDy
1	-20.0000	-19.9997
X	160.0000	159.9983
X^2Y	-10.0000	-9.9999

Table 4.1: SINDy functions identification for the species X equation (4.12). All the other terms are identified as 0.0.

Nevertheless, as we mentioned previously, in order to obtain functions for X that would resemble very much the original ones, we obtained a very distant prediction for Y. This comparison, as well as the respective fit qualities, is made in table 4.3.

$\lambda \in [1.2, 2.8]$, $\chi^2_Y = 0.112$

	Y - SINDy	Y
1	-643.7	0.0
X	14.7	-15.0
Y	538.9	0.0
X^2	0.0	0.0
XY	7.1	0.0
Y^2	-149.1	0.0
X^3	0.0	0.0
X^2Y	0.0	1.0
XY^2	-5.1	0.0
Y^3	13.7	0.0
X^4	0.0	0.0
X^3Y	0.0	0.0
X^2Y^2	0.0	0.0
XY^3	0.0	0.0
Y^4	0.0	0.0

$\lambda \in [2.9, 3.5]$, $\chi^2_Y = 0.145$

	Y - SINDy	Y
1	160.1	0.0
X	-19.2	-15.0
Y	-82.3	0.0
X^2	0.0	0.0
XY	22.9	0.0
Y^2	10.8	0.0
X^3	0.0	0.0
X^2Y	0.0	1.0
XY^2	-7.2	0.0
Y^3	0.0	0.0
X^4	0.0	0.0
X^3Y	0.0	0.0
X^2Y^2	0.0	0.0
XY^3	0.0	0.0
Y^4	0.0	0.0

$\lambda \in [3.6, 6.7]$, $\chi^2_Y = 4.511$

	Y - SINDy	Y
1	-546.0	0.0
X	33.6	-15.0
Y	277.0	0.0
X^2	0.0	0.0
XY	-8.3	0.0
Y^2	-35.2	0.0
X^3	0.0	0.0
X^2Y	0.0	1.0
XY^2	0.0	0.0
Y^3	0.0	0.0
X^4	0.0	0.0
X^3Y	0.0	0.0
X^2Y^2	0.0	0.0
XY^3	0.0	0.0
Y^4	0.0	0.0

Table 4.2: SINDy functions identification for the species Y equation (4.12) when X is correctly identified.

In this model, we had to compromise Y results in order to obtain a good prediction for X terms: we attain a good prediction for the latter species for a minimum of $\lambda = 1.2$, which is bigger than one of the parameters in Y system, 1.0.

In other respects, the good values for Y were obtained when $\lambda \in [0.1, 0.6]$, and these are detailed in table 4.3.

$\lambda \in [0.1, 0.6]$, $\chi^2_Y = 2.8 \times 10^{-6}$

	Y	Y - SINDy
X	-15.0000	-15.0000
X^2Y	1.0000	1.0000

Table 4.3: SINDy functions identification for the species Y equation (4.12).

In this λ range, the predicted terms for X are not accurate, just like it happened in the opposite situation. The description of the terms for X in these λ range is made in table 4.4.

$\lambda \in [0.1, 0.6]$, $\chi^2_Y = 0.010$

	X - SINDy	X
1	-4772.9	-20.0
X	2053.0	160.0
Y	4773.6	0.0
X^2	-0131.3	0.0
XY	-1679.2	0.0
Y^2	-1790.0	0.0
X^3	2.5	0.0
X^2Y	69.0	-10.0
XY^2	502.1	0.0
Y^3	296.5	0.0
X^4	0.0	0.0
X^3Y	-0.7	0.0
X^2Y^2	-12.0	0.0
XY^3	-50.5	0.0
Y^4	-18.3	0.0

Table 4.4: SINDy functions identification for the species X equation (4.12) when Y is correctly identified.

In summary, we obtained a correct prediction for the X species system when $\lambda \in [1.2, 6.7]$ and for Y species when $\lambda \in [0.1, 0.6]$, and we recall that we had to fit each species separately, since each time the algorithm converged for one of them the system identification was rather poor for the other. In the these ranges, the algorithm was quite successful in finding the equations of the dynamical system. In all cases, the second derivatives from the algorithm visually matched the second derivatives calculated directly from the data, and this is shown in figure 4.2.

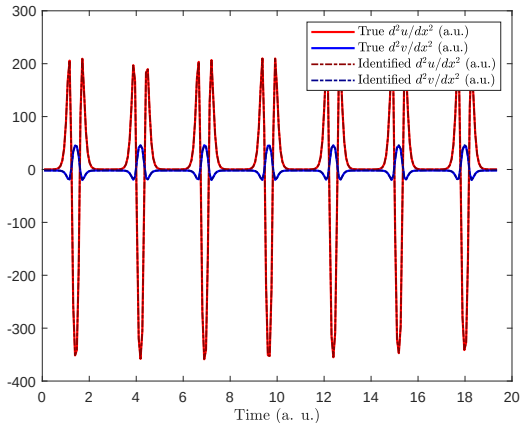


Figure 4.2: Data plot of the second derivatives computed with the SINDy output functions and the ones computed directly from the data for the steady-state of the system (4.12). Code at *Brusselator_benchmarking_1.m*.

We also verified if the first fit to the data (that is, the Moore-Penrose pseudoinverse when $\lambda = 0$) - see table 4.5 - was able to produce Turing patterns. It was not: setting an initial random profile for both species, we tested all combinations for diffusion coefficients 1.0, 0.1 and 0.01; in all these arrangements, occurred a numeric overflow almost immediately.

$\chi^2_X = 8.471 \times 10^{-6}$, $\chi^2_Y = 8.159 \times 10^{-6}$

	X	Y
1	-5.00	0.05
X	2.17	0.03
Y	4.99	0.05
X^2	-0.17	0.00
XY	-1.79	0.01
Y^2	-1.87	0.02
X^3	0.01	0.00
X^2Y	0.09	0.00
XY^2	-0.54	0.00
Y^3	0.31	0.00
X^4	-0.00	0.00
X^3Y	-0.00	0.00
X^2Y^2	-0.02	0.00
XY^3	-0.05	0.00
Y^4	-0.02	0.00

$\chi^2_X = 1.092 \times 10^{-5}$, $\chi^2_Y = 1.007 \times 10^{-5}$

	Y - SINDy	Y
1	-344.58	7.30
X	225.00	-16.89
Y	240.11	-5.38
X^2	-2.94	0.10
XY	-41.31	1.20
Y^2	-59.00	1.32
X^3	0.03	-0.00
X^2Y	-8.96	0.97
XY^2	6.86	-0.20
Y^3	4.80	-0.17

Table 4.5: Moore-Penrose inverse of Ξ for the Brusselator system (4.12) with polynomials until 4th order (left) and 3rd order (right).

4.2.1 Relationship between λ, sparsity and fit quality

As we have mentioned before, [41] proposes that the sparsifying parameter is the one which represents a compromise between the sparsity of the system and the fit quality to the data. In order

to study this hypothesis, and taking advantage of the fact that we know *a priori* what is the good λ for the SINDy algorithm in this system, we analyze how both sparsity and fit quality evolve with λ. We made this analysis until $\lambda = 80$, where all terms cancel: the results for species X are shown in figures 4.3 and 4.4, and in figures 4.5 and 4.6 for species Y. The points which correspond to λ values where the prediction of the system is correct are plotted in dark green.

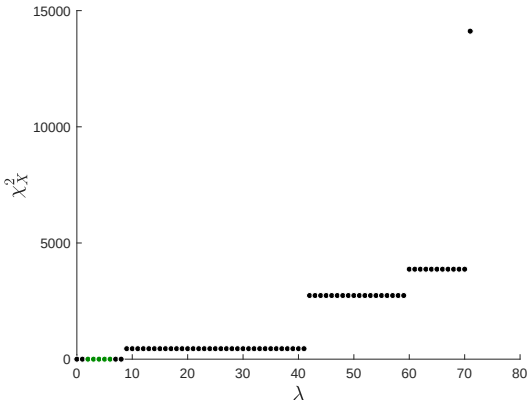


Figure 4.3: Evolution of fit quality to X species with λ.

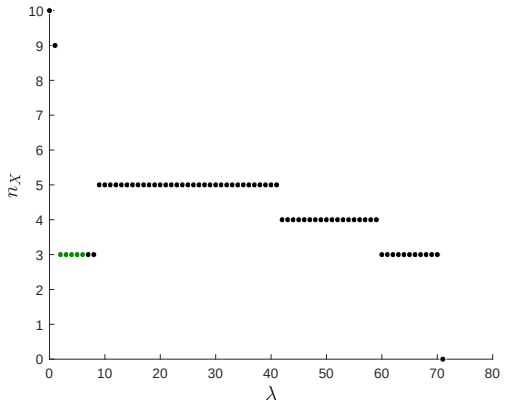


Figure 4.4: Evolution of the number of terms of species X, n_X, with λ.

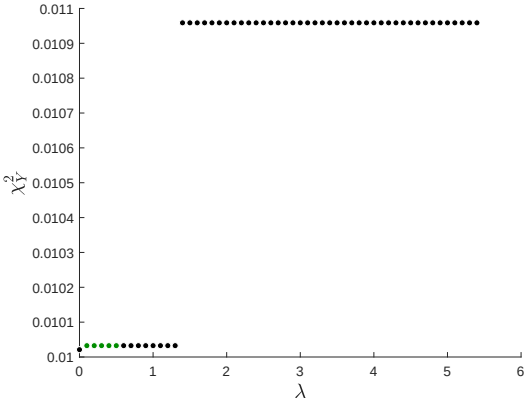


Figure 4.5: Evolution of fit quality to Y species with λ.

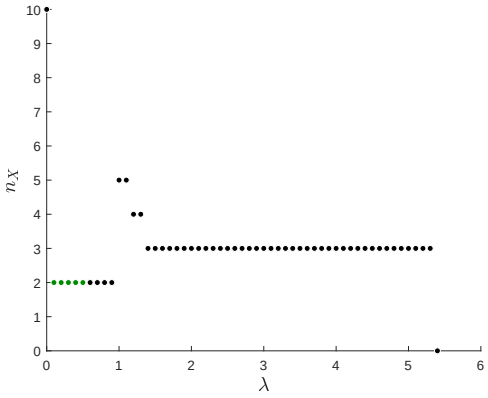


Figure 4.6: Evolution of the number of terms of species Y, n_Y, with λ.

Hence, we did not find the correct solution as a compromise between sparsity and fit quality, as it was proposed by the authors of the SINDy algorithm. This means that we need a more profound study on how to choose the sparsification parameter that correctly identifies the dynamical system.

4.3 Brusselator as a template model

Since we prospect that the pair-rule pattern is the result of activations or repressions by the space-dependent gap or maternal proteins, plus tandem regulation between them, we start by testing if the SINDy algorithm is able to identify a forced system, that is, a system with functions that have a

space dependency, as the proteins' profiles are. To do that, we are going to test the pattern which is produced by the Brusselator system when we use the parameters (3.6), but making $A = Cad(x)$ and $B = Bcd(x)$ instead. This way, the equations we want to identify are

$$\begin{aligned} \frac{d^2X(x)}{dx^2} &= -10.0\text{Bcd}(x) + 10.0X(x)\text{Cad}(x) + 10.0X(x)^2 - 10.0X(x)^2Y(x) \\ \frac{d^2Y(x)}{dx^2} &= -X(x)\text{Cad}(x) + X(x)^2Y(x), \end{aligned} \tag{4.9}$$

and this system produces a steady-state profile which is depicted in figure 4.7.

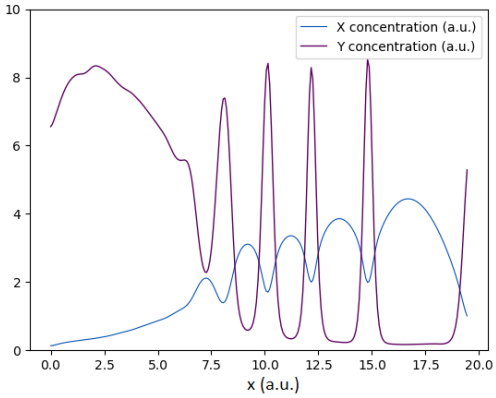


Figure 4.7: Turing pattern of the Brusselator system (4.3) with damping terms. Code at *brusselator_A_Bcd_B_Cad.py*.

Using this steady-state profile as input for the algorithm, and for candidate functions polynomials until 3rd order of $X(x)$, $Y(x)$, $Bcd(x)$ and $Cad(x)$, there is an interval of $\lambda \in [0.1, 0.9]$ where both species' equations are correctly identified, unlike what happened in the identification of the simple Brusselator system. In table 4.6, the predicted terms are compared with the original ones, and the plot of the second derivatives from the SINDy algorithm and the derivatives computed from the data are depicted in figure 4.8.

$$\lambda \in [0.1, 0.9] \text{ , } \chi^2_X = 0.008, \chi^2_Y = 2.276 \times 10^{-6}$$

	X	X - SINDy	Y	Y - SINDy
X	10.000	9.996	0.000	0.000
Bcd	-10.000	-9.996	0.000	0.000
$X\,Cad$	10.000	9.999	-1.000	-1.000
X^2Y	-10.000	-9.999	1.000	1.000

Table 4.6: SINDy functions identification for the species X and Y for the (4.3). All the other terms were correctly identified as 0.000.

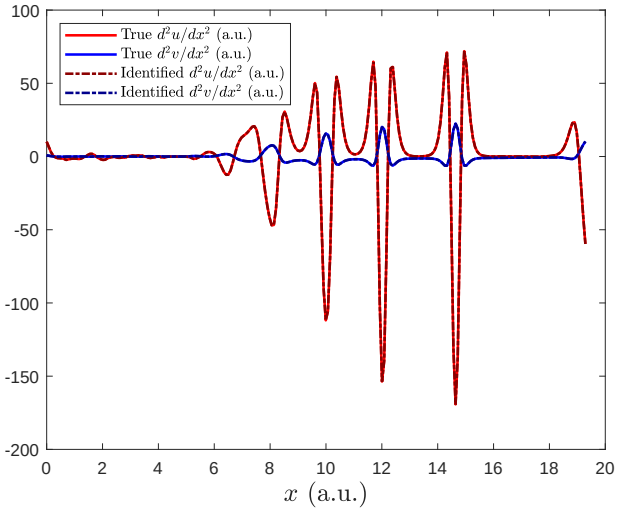


Figure 4.8: Data plot of the second derivatives computed with the SINDy output functions and the ones computed directly from the data for the steady-state of the system (4.3). Code at *Brusselator_benchmarking_bcd_cad.m*

Hence, we conclude that SINDy is able to identify the equations of a forced system. In order to identify the interaction between the pair-rule, we now fit the central part of the pattern only, which for simplicity we approximate with:

$$\begin{aligned} X(x) &= 1 + \sin(x) \\ Y(x) &= 1 + \sin(x - \pi), \end{aligned} \tag{4.10}$$

and this regulation was identified by the algorithm as

$$\begin{aligned} \frac{d^2X(x,t)}{dx^2} &= 0.125(-X^3 - X^2Y + XY^2 + Y^3) \\ \frac{d^2Y(x)}{dx^2} &= 0.125(X^3 + X^2Y - XY^2 - Y^3), \end{aligned} \tag{4.11}$$

when we used until 3^{rd} combinations of X and Y as candidate functions. These equations can also be written as

$$\begin{aligned} \frac{d^2X(x,t)}{dx^2} &= 0.125(Y^2 - X^2)(X + Y) \\ \frac{d^2Y(x)}{dx^2} &= 0.125(X^2 - Y^2)(X + Y). \end{aligned} \tag{4.12}$$

This system was identified in the interval $\lambda \in [0.00, 0.15]$, with $\chi_X = \chi_Y = 0$ and, for bigger values of λ, all terms cancel, which means that in this particular case we do not need a criterion to balance fit quality and sparsity. The comparison between the original second derivatives and the ones computed from the identified system is shown in figure 4.9.

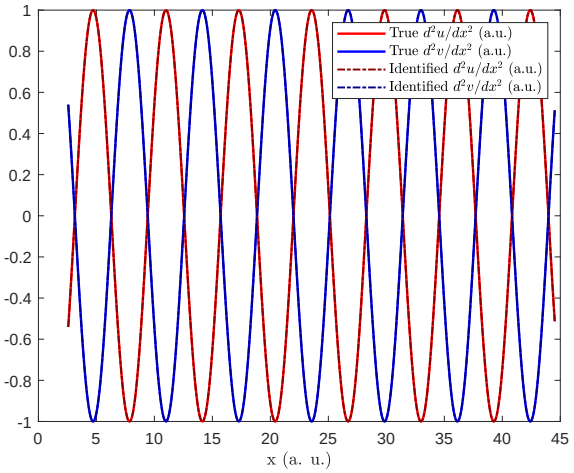


Figure 4.9: Data plot of the second derivatives computed with the SINDy output functions and the ones computed directly from the data for the steady-state of the system (4.3). Code at *sin_and_cos.m*.

Although the system's terms are clearly identified, the reaction-diffusion equations diverged with all combinations of 1.00, 0.1 and 0.01 for the diffusion coefficients, using constant profiles with noise as initial conditions as before.

This means that, at least for some systems, the convergence may depend on more specific initial profiles. Moreover, for further identification of the complete pair-rule pattern, other solutions that we lose when we make the system reduction should also be studied, which can be done by using the transient states.

5

Conclusions and Future Work

Contents

5.1 Conclusion

The genetic network of *Drosophila* is regulated in a cascade manner and the first two classes of proteins, maternal and gap, have already been analytically modeled [22], considering activation and repression interactions between the genes and proteins of the regulatory network, as well as the diffusion of these proteins. Statistical relationships between the pattern of the gap genes and the pattern of the pair-rule have also been studied, and a group of repressors and activators for each stripe has been found. Nevertheless, it has not been proven that each stripe is regulated individually, and it is physically more plausible that each protein is globally regulated by a group of repressors and a group of activators. In this sense, a model with this assumption and that could predict the pair-rule evolution and steady-state still misses.

In **chapter 2** we reviewed the model presented in [27] which translates a genetic network of activation and repression mechanisms into protein gradients, and used it to fit the steady-state of *Even-skipped* what would, in principle, enable us to find which proteins would be activators and which proteins would be repressors. This fit was rather poor, and the resulting pattern was nowhere near to produce a seven stripe pattern. For this reason, we concluded that local type reactions were not sufficient to produce such regularities, and a more robust model would be necessary.

Accordingly, in **chapter 3**, we tested if the gap and maternal proteins can regulate the 14A1 pattern of the pair-rule proteins *Eve* and *Ftz*, and this earlier stage fit was in a closer agreement with the experimental data. We proposed that these initial patterns, as they were the combination of repressing and activating proteins, were feeding the evolution of the pair-rule; moreover, we considered diffusion and an interaction between them - which means that we proposed that the seven stripes are a Turing pattern, developed via a reaction-diffusion mechanism. Based on the Brusselator, an extensively studied reaction-diffusion system with Turing patterns for a relatively flexible range of parameters, and using the initial patterns for A and B - the intermediary species of the Brusselator - we were able to obtain steady-state and transient profiles which were in a very good agreement with the experimental data. Different number of stripes were obtained for different values of the diffusion coefficients, and seven stripes patterns were achieved when $D_X \in [0.04D_Y, 0.06D_Y]$.

We also extended this approach for the segment-polarity proteins. The biological experiments indicate that there is one segment-polarity, *en*, which is activated by both *Eve* and *Ftz*, and another segment-polarity, *wg*, which is, in contrary, repressed by the same two pair-rule. Once more, beginning with the reaction model for these regulations, we were able to outline the initial patterns for these proteins. From these early patterns on, and considering the Brusselator system for the reaction-diffusion mechanism, we were able to obtain the fourteen stripe pattern when $D_X = 0.03D_Y$.

Finally, in **chapter 4**, we aimed to identify the dynamical system which regulates the interaction between the pair-rule or between the segment-polarity. To do this, we took advantage of the regression algorithm SINDy which, from steady-state profiles, is able to identify the nonlinear functions that compose the system, balancing fit quality and the system sparsity (i.e., avoiding overfitting). This algorithm was able to correctly identify the Brusselator system as well as a modification with space-dependent

functions, and the interaction terms for the pair-rules and segment-polarity were then successfully identified. Nonetheless, this system diverged when integrated, for all combinations of 1.0, 0.1 and 0.01 for the diffusion coefficients.

5.1.1 Achievements

Here, we briefly summarize the major achievements of this work. Firstly, we showed that, despite what is suggested in most literature, the local regulation via activation and repression by the earlier proteins, gap and maternal, is not sufficient to produce a high frequency pattern as the seven stripes pattern is. Moving to our second hypothesis of protein formation, the reaction-diffusion hypothesis, we simulated the development and setting of the pair-rule pattern, as well as the segment-polarity steady-state pattern, using as a template model a very established system for Turing patterns, the Brusselator. Our simulations were in a very good agreement with the experimental data, as well as with the regulatory mechanisms proposed in the literature. Using a regression algorithm for nonlinear systems, the proper interaction terms which generate the central part of the pattern were also identified, but the system was not integrated.

5.2 Future work

In order to make a deeper analysis of the results found, we suggest a series of experiments and studies to be made in the future:

- The fit of the *Ftz* initial pattern was in a very good agreement with the experimental data, and a group of repressors and activators was identified. Since the majority of biological experiments concerning *Ftz* were made with pair-rule proteins, we suggest to observe the resulting pattern of this protein with mutations in the maternal and gap, in order to see if the repressors and activators were correctly identified. The identification of the *Eve* pattern was not as good as the *Ftz*'s, in part due to a shadow defect in the data collection, which we also suggest to be corrected in future experiments.

- According to this reaction-diffusion hypothesis, the stripes are only formed when the two pair-rule species interact with each other. Nonetheless, it is observed experimentally that a mutation in *Ftz* does not affect substantially the pattern of *Eve* and vice-versa, which is not in complete agreement with our model. Due to this, we suggest further experiments considering simultaneous mutations in several pair-rule, to observe if the interaction mechanisms happen not between two species but between more, and this suggestion applies to the segment-polarity as well.

- To analyze if the proteins' diffusion coefficients found by our model are realistic or not, these constants need to be measured experimentally.

- We could not compare the evolution of our simulation with the transient profiles of the segment polarity, and we suggest to analyze this data.

- Although the regulatory terms for the pair-rule/segment-polarity were identified, the system did not converge into a Turing pattern, therefore, the solution space of the algorithm, as well as their dependence on the initial profile, need a more complete characterization. Ideally, if we had a criterion to determine whether a reaction-diffusion system is able or not to produce Turing patterns, this test could be immediately made on the SINDy solutions and no numerical integration would be necessary.

Bibliography

[1] F. Alves, *A mathematical model for segment formation in Drosophila melanogaster*. PhD thesis, Instituto Superior Técnico, 2006.

[2] C. Nüsslein-Volhard, *Coming to life - how genes drive development*. Yale University Press, 2006.

[3] S. F. Gilbert, *Developmental Biology, 6th edtion*. Sinauer Associates, 2000.

[4] A. M. Arias, N. E. Baker, and P. W. Ingham, "Role of segment polarity genes in the definition and maintenance of cell states in the drosophila embryo," *The Company of Biologists*, vol. 103, pp. 157–170, 1988.

[5] R. Dilão, "Mathematical models of morphogenesis," *ITM Web of Conferences*, vol. 4, p. 01001, 2015. doi: 10.1051/itmconf/2015040100.

[6] S. Parkhurst and D. Ish-Horowicz, "Mis-regulating segmentation gene expression in *Drosophila*," *Development*, vol. 111, pp. 1121–1135, 1991.

[7] P. W. Ingham, "The molecular genetics of embryonic pattern formation in *Drosophila*," *Nature*, vol. 335, p. 744, 1988.

[8] S. Small, A. Blair, and M. Levine, "Regulation of two pair-rule stripes by a single enhancer in the drosophila embryo," *Developmental Biology*, vol. 175, pp. 314–324, 1996.

[9] D. Stanojevic, S. Small, and M. Levine, "Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo," *Science*, vol. 254, pp. 1385–7, 1991.

[10] S. Small, R. Kraut, T. Hoey, R. Warrior, and M. Levine, "Transcriptional regulation of a pair-rule stripe in *Drosophila*.," *Genes Dev*, vol. 5, pp. 827–39, 1991.

[11] S. Surkova, E. Golubkova, Manu, L. Panok, L. Mamon, J. Reinitz, and M. Samsonova, "Quantitative dynamics and increased variability of segmentation gene expression in the drosophila krüppel and knirps mutants," *Developmental biology*, vol. 378, pp. 99–112, 2013.

[12] M. Frasch and M. Levine, "Complementary pattern of *even-skipped* and *fushi-tarazu* expression involve their differential regulation by a common set of segmentation genes in *Drosophila*," *Genes & Development*, vol. 1, pp. 981–995, 1987.

[13] K. Harding, T. Hoey, R. Warrior, and M. Levine, "Autoregulatory and gap gene response elements of the even-skipped promoter of drosophila," *The EMBO journal*, vol. 8, pp. 1205–12, 1989.

[14] J. Jiang, T. Hoey, and M. Levine, “Autoregulation of a segmentation gene in drosophila: combinatorial interaction of the even-skipped homeo box protein with a distal enhancer element.,” *Genes Dev*, vol. 5, pp. 265–277, 1991.

[15] K. Harding, C. Rushlow, H. J. Doyle, T. Hoey, and M.Levine, “Cross-regulatory interactions among pair-rule genes in *Drosophila*,” *Science*, vol. 233, pp. 953–959, 1986.

[16] Y. Yu and L. Pick, “Non-periodic cues generate seven *ftz* stripes in the *Drosophila* embryo,” *Mechanisms of Development*, vol. 50, pp. 163–175, 1995.

[17] P. W. Ingham and A. Martinez-Arias, “The correct activation of *Antennapedia* and bithorax complex genes requires the *fushi tarazu* gene,” *Nature*, vol. 324, p. 592–597, 1986.

[18] C. Tsai and P. Gergen, “Pair-rule expression of the drosophila fushi tarazu gene: A nuclear receptor response element mediates the opposing regulatory effects of runt and hairy,” *Development*, vol. 191, pp. 453–462, 1995.

[19] J. E. Butler and J. T. Kadonaga, “The rna polymerase ii core promoter: a key component in the regulation of gene expression,” *Genes & Development*, vol. 16, p. 2583–2592, 2002.

[20] Y. Hiromi and W. J. Gehring, “Regulation and function of the drosophila segmentation gene *fushi tarazu*,” *Cell*, vol. 50, pp. 963–934, 1987.

[21] R. Dilão and D. Muraro, “Calibration and validation of a genetic regulatory network model describing the production of the protein hunchback in drosophila early development,” *Comptes Rendus Biologies*, vol. 333, pp. 779–788, 2010.

[22] F. Alves and R. Dilão, “Modeling segmental patterning in *Drosophila*: Maternal and gap genes,” *Journal of Theoretical Biology*, vol. 241, pp. 342–359, 2006.

[23] M. Samee and S. Sinha, “Quantitative modeling of a gene’s expression from its intergenic sequence,” *PLoS Computational Biology*, vol. 10, 2014.

[24] G. R. Ilsley, J. Fisher, R. Apweiler, A. H. DePace, and N. M. Luscombe, “Cellular resolution models for even skipped regulation in the entire drosophila embryo,” *PLoS Computational Biology*, vol. 2, 2013.

[25] D. Stanojevic, T. Hoey, and M. Levine, “Sequence-specific dna-binding activities of the gap proteins encoded by *hunchback* and *Kruppel* in *Drosophila*,” *Nature*, vol. 341, pp. 331–335, 1989.

[26] F. Crick, “Central dogma of molecular biology,” *Nature*, vol. 227, pp. 561–563, 1970.

[27] F. Alves and R. Dilão, “A simple framework to describe the regulation of gene expression in prokaryotes,” *Comptes Rendus Biologies*, vol. 328, pp. 429–44, 2005.

[28] “Gene Expression.” `https://www.nature.com/scitable/topicpage/gene-expression-14121669/`, 2014.

[29] S. K. S., *Chemical Chaos*. Oxford University Press, 1993.

[30] R. Dilão, "The reaction-diffusion approach to morphogenesis," *Proceedings of 4th Brazilian Symposium on Mathematical and Computational Biology*, vol. 1, 2004.

[31] R. Dilão and D. Muraro, "A software tool to model genetic regulatory networks," *PLoS ONE*, vol. 5, pp. 1–10, 2010.

[32] A. M. Turing, "The chemical basis of morphogenesis," *Philosophical Transactions of the Royal Society of London*, vol. 237, pp. 5–72, 1952.

[33] D. E. Strier and S. P. Dawson, "Turing patterns inside cells," *PLoS ONE*, vol. 2, 2007.

[34] T. W. Hiscock and S. G. Megason, "Mathematically guided approaches to distinguish models of periodic patterning," *The Company of Biologists*, vol. 142, pp. 409–419, 2015.

[35] R. Dilão and J. Sainhas, "Validation and calibration of models for reaction-diffusion systems," *International Journal of Bifurcations and Chaos in Applied Sciences and Engineering*, vol. 8, pp. 1163–1182, 1998.

[36] K. Howard, "The generation of periodic patterns during early *Drosophila* embryogenesis," *The Company of Biologists Limited*, vol. 104, pp. 35–50, 1988.

[37] S. Kondo and T. Miura, "Reaction-diffusion model as a framework for understanding biological pattern formation," *Science*, vol. 329, pp. 1616–1620, 2012.

[38] H. Meinhardt, "Hierarchical inductions of cell states: a model for segmentation in *Drosophila*," *Journal of Cell Science*, vol. 4, pp. 357–381, 1986.

[39] J. M. W. Slack, *Essential Developmental Biology*. Wiley-Blackwell, 2001.

[40] S. F. Gilbert, *Developmental Biology*. Sinauer Associates, 2000.

[41] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *PNAS*, vol. 113, pp. 3932–3937, 2016.